COMPUTATIONAL NEUROIMAGE ANALYSIS TOOLS FOR BRAIN (DISEASES) BIOMARKERS

EDITED BY: Diana M. Sima, Meritxell Bach Cuadra, Tim B. Dyrby and Koen Van Leemput

PUBLISHED IN: Frontiers in Neuroscience and Frontiers in Neurology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-344-5
DOI 10.3389/978-2-88974-344-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL NEUROIMAGE ANALYSIS TOOLS FOR BRAIN (DISEASES) BIOMARKERS

Topic Editors:
Diana M. Sima, Icometrix (Belgium), Belgium
Meritxell Bach Cuadra, Lausanne University Hospital (CHUV/UNIL), Switzerland
Tim B. Dyrby, Technical University of Denmark, Denmark
Koen Van Leemput, Harvard Medical School, United States

Citation: Sima, D. M., Cuadra, M. B., Dyrby, T. B., Van Leemput, K., eds. (2022). Computational Neuroimage Analysis Tools for Brain (Diseases) Biomarkers. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-344-5

Table of Contents

05 ***Editorial: Computational Neuroimage Analysis Tools for Brain (Diseases) Biomarkers***
Diana M. Sima, Meritxell Bach Cuadra, Tim B. Dyrby and Koen Van Leemput

08 ***Multiplex Networks to Characterize Seizure Development in Traumatic Brain Injury Patients***
Marianna La Rocca, Rachael Garner, Nicola Amoroso, Evan S. Lutkenhoff, Martin M. Monti, Paul Vespa, Arthur W. Toga and Dominique Duncan

19 ***RatLesNetv2: A Fully Convolutional Network for Rodent Brain Lesion Segmentation***
Juan Miguel Valverde, Artem Shatillo, Riccardo De Feo, Olli Gröhn, Alejandra Sierra and Jussi Tohka

30 ***Diagnostic Performance of MRI Volumetry in Epilepsy Patients With Hippocampal Sclerosis Supported Through a Random Forest Automatic Classification Algorithm***
Juan Pablo Princich, Patricio Andres Donnelly-Kehoe, Alvaro Deleglise, Mariana Nahir Vallejo-Azar, Guido Orlando Pascariello, Pablo Seoane, Jose Gabriel Veron Do Santos, Santiago Collavini, Alejandro Hugo Nasimbera and Silvia Kochen

46 ***Diffusion Tensor Imaging Group Analysis Using Tract Profiling and Directional Statistics***
Mehmet Özer Metin and Didem Gökçay

63 ***GAMER-MRI in Multiple Sclerosis Identifies the Diffusion-Based Microstructural Measures That Are Most Sensitive to Focal Damage: A Deep-Learning-Based Analysis and Clinico-Biological Validation***
Po-Jui Lu, Muhamed Barakovic, Matthias Weigel, Reza Rahmanzadeh, Riccardo Galbusera, Simona Schiavi, Alessandro Daducci, Francesco La Rosa, Meritxell Bach Cuadra, Robin Sandkühler, Jens Kuhle, Ludwig Kappos, Philippe Cattin and Cristina Granziera

73 ***Management and Quality Control of Large Neuroimaging Datasets: Developments From the Barcelona*β*eta Brain Research Center***
Jordi Huguet, Carles Falcon, David Fusté, Sergi Girona, David Vicente, José Luis Molinuevo, Juan Domingo Gispert, and Grégory Operto for the ALFA Study

88 ***Transductive Transfer Learning for Domain Adaptation in Brain Magnetic Resonance Image Segmentation***
Kaisar Kushibar, Mostafa Salem, Sergi Valverde, Àlex Rovira, Joaquim Salvi, Arnau Oliver and Xavier Lladó

103 ***Harmonized Segmentation of Neonatal Brain MRI***
Irina Grigorescu, Lucy Vanes, Alena Uus, Dafnis Batalle, Lucilio Cordero-Grande, Chiara Nosarti, A. David Edwards, Joseph V. Hajnal, Marc Modat and Maria Deprez

120 ***Bundle-Specific Axon Diameter Index as a New Contrast to Differentiate White Matter Tracts***
Muhamed Barakovic, Gabriel Girard, Simona Schiavi, David Romascano, Maxime Descoteaux, Cristina Granziera, Derek K. Jones, Giorgio M. Innocenti, Jean-Philippe Thiran and Alessandro Daducci

133 ***Multi-Channel 4D Parametrized Atlas of Macro- and Microstructural Neonatal Brain Development***
Alena Uus, Irina Grigorescu, Maximilian Pietsch, Dafnis Batalle, Daan Christiaens, Emer Hughes, Jana Hutter, Lucilio Cordero Grande, Anthony N. Price, Jacques-Donald Tournier, Mary A. Rutherford, Serena J. Counsell, Joseph V. Hajnal, A. David Edwards and Maria Deprez

154 ***Fast and High-Resolution Neonatal Brain MRI Through Super-Resolution Reconstruction From Acquisitions With Variable Slice Selection Direction***
Yao Sui, Onur Afacan, Ali Gholipour and Simon K. Warfield

169 ***The Influence of Radio-Frequency Transmit Field Inhomogeneities on the Accuracy of G-ratio Weighted Imaging***
Tim M. Emmenegger, Gergely David, Mohammad Ashtarayeh, Francisco J. Fritz, Isabel Ellerbrock, Gunther Helms, Evelyne Balteau, Patrick Freund and Siawoosh Mohammadi

186 ***A Contrast Augmentation Approach to Improve Multi-Scanner Generalization in MRI***
Maria Ines Meyer, Ezequiel de la Rosa, Nuno Pedrosa de Barros, Roberto Paolella, Koen Van Leemput and Diana M. Sima

201 ***Altered Spontaneous Brain Activity Related to Neurologic and Sleep Dysfunction in Children With Obstructive Sleep Apnea Syndrome***
Jie Bai, Hongwei Wen, Jun Tai, Yun Peng, Hongbin Li, Lin Mei, Tingting Ji, Xiaodan Li, Yanhua Li, Xin Ni and Yue Liu

frontiers in Neuroscience

EDITORIAL
published: 18 February 2022
doi: 10.3389/fnins.2022.841807

Editorial: Computational Neuroimage Analysis Tools for Brain (Diseases) Biomarkers

Diana M. Sima[1,2*], *Meritxell Bach Cuadra*[3,4,5], *Tim B. Dyrby*[6,7] *and Koen Van Leemput*[8,9]

[1] icometrix, Leuven, Belgium, [2] AI Supported Modelling in Clinical Sciences (AIMS), Vrije Universiteit Brussel, Brussels, Belgium, [3] Center for Biomedical Imaging (CIBM), Lausanne, Switzerland, [4] Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland, [5] Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [6] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark, [7] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Amager and Hvidovre, Copenhagen, Denmark, [8] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark, [9] Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States

Keywords: neuroimaging, biomarkers, brain disease, neurodevelopment, deep learning, domain adaptation, brain segmentation, brain connectivity

Editorial on the Research Topic

Computational Neuroimage Analysis Tools for Brain (Diseases) Biomarkers

This Research Topic focuses on recent advances in the field of neuroimaging biomarkers, including novel imaging technologies, image processing, and artificial intelligence approaches, which are pushing forward the achievement of precision medicine. Indeed, brain imaging often visualizes disease effects with greater sensitivity than clinical observation, thus holding great promise to help diagnose patients at the earliest stages of their disease, when treatment is most effective.

To unleash this potential, neuroimaging biomarkers should be proven to be useful, sensitive and reliable. Nowadays, conventional magnetic resonance imaging is complemented by numerous advanced acquisition and processing techniques, aiming to unravel structural and functional brain connectivity and pathological alterations in brain tissue up to the microstructural level.

Current challenges amount not only to designing clinically feasible acquisition protocols and reliable image processing methods, but also integrating the wealth of data that gets collected in different centers and in different neuroimaging domains in a consistent way to permit reuse in new domains. As large-scale and more complex neuroimaging datasets have been or are being collected in heterogeneous ways by various organizations (with the field of Alzheimer's disease as a notable example), Huguet et al. propose core principles to facilitate reusability and data sharing. They implement an ecosystem of modules and tools, including automated quality control, which is suitable for large neuroimaging studies.

However, the benefits of merging different datasets is often counteracted by their heterogeneous nature. Indeed, systematic differences can occur due to site-specific conditions or due to bias in population characteristics. Moreover, some tasks are difficult to generalize from one dataset to another, since not all datasets are consistently labeled, or are labeled at all. As such, Kushibar et al. propose a transductive transfer learning approach for domain adaptation to reduce the domain-shift effect in human brain MRI segmentation. The transductive property means that there are two disjoint source and target domains, where label annotations are only available in the source domain, but examples from the target domain are also present. The proposed network is jointly optimized by integrating both source and target images into the transductive training process, minimizing

OPEN ACCESS

Edited and reviewed by:
Vince D. Calhoun,
Georgia State University,
United States

***Correspondence:**
Diana M. Sima
diana.sima@icometrix.com

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *22 December 2021*
Accepted: *26 January 2022*
Published: *18 February 2022*

Citation:
Sima DM, Bach Cuadra M, Dyrby TB and Van Leemput K (2022) Editorial: Computational Neuroimage Analysis Tools for Brain (Diseases) Biomarkers. Front. Neurosci. 16:841807. doi: 10.3389/fnins.2022.841807

the domain-shift effect with a histogram loss at feature level. The method shows performance improvements up to 10% in terms of Dice similarity coefficient for the segmentation of subcortical brain structures and white matter hyperintensities, compared to a model pre-trained in the source domain only.

Meyer et al. propose to increase the generalization capability of state-of-the-art convolutional neural network (CNN) models trained on homogeneous datasets by applying an intensity-based data augmentation approach based on Gaussian mixture modeling. This approach is shown to significantly impact the generalization performance of brain structures segmentation when the training set is very homogeneous, but also when it consists of heterogeneous multi-scanner brain images.

Valverde et al. employ CNNs for the task of lesion segmentation in rodent brains. They suggest that an architecture resembling an autoencoder has better performance compared with three other convolutional neural networks specifically designed for medical image segmentation. Moreover, when comparing versions trained on homogeneous and heterogeneous training datasets, it is clear that increasing training data diversity improves performance, as measured by the capability to extrapolate to different-looking ischemic brain lesions.

Not only combining datasets, but also combining existing modeling methods can be beneficial for obtaining more robust biomarkers. This is illustrated by Lu et al., who focus on quantitative diffusion measures for classifying multiple sclerosis lesions. They use eight open-source biophysical models of multishell diffusion data to reconstruct the isotropic and intra-axonal compartments, and identify the microstructural diffusion measures that are most discriminative for focal pathology. Further, they show that some of the combinations of the selected normalized diffusion measures better correlate with patients' disability and neuroaxonal damage than individual measures.

While deep learning techniques play an important role in current neuroimaging research due to their ability to increase reliability of computed biomarkers, there are still essential gains to be made by improving MRI acquisitions and parameter map reconstructions. For instance, Emmenegger et al. assess the effect of radio-frequency transmit (B1+) field inhomogeneities correction on the accuracy of MR G-ratio weighted imaging, which is an aggregated measure of relative myelination of axons across the entire brain white matter. B1+ correction *via* a measured B1+ field map is the method of choice to reduce bias and test-retest error. However, if the B1+ field map cannot be acquired, a data-driven B1+ correction approach is also proposed, and shown to reduce the error and bias by a factor of three.

Metin and Gökçay highlight the value of using directional information from diffusion tensor imaging of the brain for group statistics, rather than scalar metrics that consider only the magnitude of the diffusion. A typical scalar metric used in group studies is the Fractional Anisotropy map. Directional statistical analysis is particularly important along the white matter tracts, especially when the tract length increases.

Barakovic et al. propose a more robust estimation of the axon diameter index of pathways by jointly estimating the microstructure properties of the tissue and the macroscopic organization of the white matter connectivity. The method overcomes limitations of previous voxel-wise approaches, which neglect the fact that axons are continuous three-dimensional structures that are not limited to the extent of each voxel. By computing the axon diameter index in bundles of streamlines, where each streamline represents a group of axons that share a similar trajectory, the method is able to estimate an average diameter for the represented group of axons. As such, they show that the fiber bundle composition agrees with histology and known anatomy.

The value of brain imaging biomarkers is crucial since early stages of development. Thus, imaging the neonatal brain with the aim of establishing patterns of (normal or abnormal) brain development is also an important and very active topic of research. It is now possible to acquire high resolution (isotropic 0.4 mm) images in a short time (6 min), due to novel super-resolution reconstruction of three short duration scans with variable directions of slice selection (Sui et al.).

Uus et al. highlight the importance of employing a suitable atlas of normal neonatal brain development. They build a single multi-channel spatio-temporal atlas based on multiple metrics extracted from both diffusion and structural MRI, and then compare, in this novel atlas space, two groups of neonates: born at term and preterm. Significant effects linked to prematurity are shown to be present in multiple brain regions, including the transient fetal compartments, indicating that white matter maturation is altered by preterm birth.

Grigorescu et al. predict tissue segmentation maps of neonates on T2-weighted magnetic resonance imaging data. Similarly to Kushibar et al., they employ domain adaptation techniques for the challenging task of brain segmentation in a preterm-born neonatal population, where the training set consists of an annotated dataset acquired from term-born neonates with a different scanner and acquisition protocol. Importantly, adding the domain adaptation to the model did not degrade performance in the source domain. Moreover, in line with Meyer et al., the authors show the importance of adding data augmentation during training.

Finally, neuroimaging methods should not only act as tools for new scientific discoveries, but should also provide practical solutions for different neurological conditions in clinical practice. As such, La Rocca et al. develop a novel approach based on structural MRI to analyse interactions between brain components. Based on these novel features, they train a classifier able to predict with an accuracy of 70% whether subjects who suffered a mild traumatic brain injury will have at least one seizure in the future. Princich et al. revisit known MRI volumetry biomarkers that can assist in the diagnosis of temporal lobe epilepsy. Despite observing differences between different MRI segmentation software in terms of hippocampal volumetry (here, FreeSurfer and volBrain are compared), the study strongly reinforces the value of hippocampal asymmetry in differentiating healthy controls and epilepsy patients with hippocampal sclerosis.

Bai et al. focus on childhood obstructive sleep apnea, a sleep-related breathing disorder that can have an important negative impact on neurological development. Functional MRI reveals altered spontaneous brain activity, with dysfunctions occurring in the default mode network, the frontal lobe, and the lingual gyrus.

In conclusion, this Research Topic clearly illustrates that the field of computational neuroimaging is active and fascinating, with a wide range of novel methodologies aiming at reliability and generalizability through domain adaptation, data augmentation, super-resolution, and quality control. Important applications presented here aim at understanding brain development, connectivity and microstructure, as well as brain diseases such as multiple sclerosis, epilepsy, post-traumatic brain injury, and sleep dysfunction.

AUTHOR CONTRIBUTIONS

DMS wrote the draft of the manuscript. MBC, TBD, and KVL contributed to manuscript revision, read, and approved the submitted version. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to acknowledge the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 765148 (TRABIT).

Conflict of Interest: DMS was employed by icometrix.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 30 November 2020
doi: 10.3389/fnins.2020.591662

Multiplex Networks to Characterize Seizure Development in Traumatic Brain Injury Patients

Marianna La Rocca[1], Rachael Garner[1], Nicola Amoroso[2], Evan S. Lutkenhoff[3], Martin M. Monti[3], Paul Vespa[4], Arthur W. Toga[1] and Dominique Duncan[1]*

[1] Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States, [2] Dipartimento di Farmacia - Scienze del Farmaco, Università degli Studi di Bari "A. Moro", Bari, Italy, [3] Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States, [4] David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, United States

Traumatic brain injury (TBI) may cause secondary debilitating problems, such as post-traumatic epilepsy (PTE), which occurs with unprovoked recurrent seizures, months or even years after TBI. Currently, the Epilepsy Bioinformatics Study for Antiepileptogenic Therapy (EpiBioS4Rx) has been enrolling moderate-severe TBI patients with the goal to identify biomarkers of epileptogenesis that may help to prevent seizure occurrence and better understand the mechanism underlying PTE. In this work, we used a novel complex network approach based on segmenting T1-weighted Magnetic Resonance Imaging (MRI) scans in patches of the same dimension (network nodes) and measured pairwise patch similarities using Pearson's correlation (network connections). This network model allowed us to obtain a series of single and multiplex network metrics to comprehensively analyze the different interactions between brain components and capture structural MRI alterations related to seizure development. We used these complex network features to train a Random Forest (RF) classifier and predict, with an accuracy of 70 and a 95% confidence interval of [67, 73%], which subjects from EpiBioS4Rx have had at least one seizure after a TBI. This complex network approach also allowed the identification of the most informative scales and brain areas for the discrimination between the two clinical groups: seizure-free and seizure-affected subjects, demonstrating to be a promising pilot study which, in the future, may serve to identify and validate biomarkers of PTE.

Keywords: post-traumatic epilepsy, traumatic brain injury, structural magnetic resonance imaging, multiplex networks, random forest, machine learning, complex networks

OPEN ACCESS

Edited by:
Diana M. Sima, Icometrix, Belgium

Reviewed by:
Baxter P. Rogers, Vanderbilt University, United States
Pierre Besson, Northwestern University, United States

****Correspondence:***
Marianna La Rocca
marianna.larocca@loni.usc.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *05 August 2020*
Accepted: *09 November 2020*
Published: *30 November 2020*

Citation:
La Rocca M, Garner R, Amoroso N, Lutkenhoff ES, Monti MM, Vespa P, Toga AW and Duncan D (2020) Multiplex Networks to Characterize Seizure Development in Traumatic Brain Injury Patients. Front. Neurosci. 14:591662. doi: 10.3389/fnins.2020.591662

1. INTRODUCTION

Traumatic brain injury (TBI) is the third most common cause of death and debilitating secondary problems in adults and children worldwide. One common consequence of TBI that causes significant disability amongst patient populations is post-traumatic epilepsy (PTE) (Humphreys et al., 2013). This condition develops in up to 50% of patients with TBI. Post-traumatic epilepsy (PTE) is diagnosed if two or more unprovoked seizures occur at least 1 week after a TBI (Diaz-Arrastia et al., 2009). Recent investigations suggest that injury severity and especially epileptic activity are high risk factors of PTE, although the mechanisms by which trauma to the brain

tissue leads to recurrent seizures is not known. Therefore, studying if specific structural Magnetic Resonance Imaging (sMRI) changes can be related to seizures after a TBI is of fundamental importance to carry out the first steps toward the discovery of early biomarkers of PTE (Kim et al., 2018). PTE is not a homogeneous condition and can appear weeks or several years after a TBI. As a consequence, the precise percentage of TBI patients who develop PTE is not known (Verellen and Cavazos, 2010). Currently, growing attention has been devoted to investigate PTE. In this regard, the Epilepsy Bioinformatics Study for Antiepileptogenic Therapy (EpiBioS4Rx) is an international, multi-center project conceived to identify biomarkers of epileptogenesis after a TBI in order to evaluate treatments that could prevent the development of PTE and design clinical trials of antiepileptogenic therapies on an extensive patient population. With this project, the scientific community can be granted access to a large amount of high quality, multi-modal data, including imaging, electrophysiology, and clinical data from both humans and animals.

Changes in gray matter and white matter related to epilepsy have been widely observed by using structural MRI (Immonen et al., 2018; Shah et al., 2019; Lutkenhoff et al., 2020). Many recent studies have shown that machine learning techniques and multiplex networks applied to completely non-invasive neuroimaging techniques, such as structural MRI, can be useful and efficient to detect pathological alterations in several neurological diseases, such as Alzheimer's disease, Parkinson's disease, and epilepsy (Amoroso et al., 2018c; La Rocca et al., 2018; Bharath et al., 2019). Multiplex networks overcome the limit of the existing complex network standard approaches not to be able to collectively study what happens to the same nodes as their interactions change. In our previous work (Garner et al., 2019), we used different machine learning strategies to identify alterations in functional brain connectivity that are related to seizure outcome following TBI. However, the present study is the first which uses the combination of multiplex networks of structural MRIs and machine learning techniques to distinguish patients who have developed at least one seizure after a TBI from those who have not experienced any seizures. This study is of paramount importance, because it offers an opportunity to observe alterations in TBI brain networks that may reflect structural MRI changes related to seizure development.

This paper provides three main results: (i) the implementation of a pipeline which combines complex network and machine learning models for the identification of TBI patients who have developed epilepsy; (ii) the investigation of the most appropriate scale or patch size to study seizure development in TBI patients; (iii) the implementation, on a TBI cohort, of a promising complex network model based on segmenting the brain in patches to obtain comprehensive clinical information on the whole brain. In the future, this pilot study may help clinicians localize the epileptogenic focus more precisely, relate brain lesions to seizure occurrence and understand the relationship between neuronal activity abnormalities and structural damage.

TABLE 1 | Imaging findings are reported for each clinical class.

Injury type	Seizure-free patients	Patients with seizure
Skull fracture	27/37	16/16
Epidural hematoma	8/37	4/16
Extraaxial hematoma	18/37	9/16
Acute subdural hematoma	27/37	14/16
Subarachnoid hemorrhage	30/37	14/16
Intracerebral/Intraparenchymal hemorrhage	22/37	11/16
Midline shift (Avg shift)	21/37(4.53)	8/16(6.47)
Cisternal compression	6/37	3/16
Frontal contusion	22/37	9/16
Temporal contusion	19/37	7/16
Brain edema	14/37	7/16
Penetrating injury	1/37	0/16

Injury characteristics were reported by clinical staff based on patient Computed Tomography (CT) scans on the day of hospital admission. No statistically significant between-group differences were found in the imaging findings except in the skull fracture ($p = 0.02$).

TABLE 2 | Sample size, gender, and Glasgow Coma Score (GCS) information are reported for each clinical class.

Clinical status	Sample size	Age	Female/Male	GCS score
Seizure-free patients	37	36.28 ± 21.18	4/33	10.78 ± 4.05
Patients with seizure	16	40.50 ± 18.05	3/13	8.94 ± 3.59

Age and GCS were provided in terms of mean and standard deviation. No statistically significant differences between the two classes were found with respect to age, GCS score, and gender. Statistical evaluations were performed with a Kruskal–Wallis statistic test except for the gender, for which a Chi-square test was used.

2. MATERIALS AND METHODS

2.1. Dataset

In this work, we used 53 structural MRI scans of TBI subjects recruited in EpiBioS4Rx according to specific inclusion and exclusion available online[1]. Sixteen of these subjects have experienced at least one seizure within 6 months of a TBI and 37 have not experienced any seizures. As part of their clinical care, 14 subjects required a craniectomy (3 seizure and 11 non-seizure, $p > 0.05$). Additional clinical and demographic information are reported in **Tables 1**, **2**. 3D T1-weighted volumes were acquired within 32 days (median 8 and interquartile range of [2, 15]) after the TBI using 3T Siemens, Philips, and GE scanners according to a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence with the following parameters: 256 mm field of view (FOV); 1 mm slice thickness; 1,500–2,500 ms repetition time (TR); minimum echo time (TE); 1,100–1,500 ms inversion time (TI); 8–15 degree flip angle; 256 phase-encoding steps, number of excitations (NEX) >1 and 256 Hz frequency.

[1] https://sites.google.com/g.ucla.edu/epibios4rxmobilewebsite/inclusionexclusion-criteria

2.2. MRI Processing

Protocol compliance and quality control (QC) were undertaken using the Laboratory Of Neuro Imaging (LONI) QC System[2]. Structural MRI scans were processed with the Oxford FMRIB Software Library (FSL) (Jenkinson et al., 2012). Firstly, image skull-stripping was obtained with the optimized brain extraction script for patient brain (optiBET) (Lutkenhoff et al., 2014). For those few cases where the automatic brain extraction was particularly challenging due to the significant brain deformation caused by the trauma, we performed the skull stripping by manually adjusting the brain extraction threshold with FSL Brain Extraction Tool (BET). Then MRI scan intensity differences, yielded by bias field, were normalized. After intensity normalization and brain extraction, a spatial normalization was performed to co-register the different images into a common coordinate space by using an affine transformation. The MNI152 was adopted as the reference template, and registration was performed with the FSL Linear Registration Tool (FLIRT) with a standard parameter configuration. Even though a deformable registration would have given a better overlap among TBI subjects, we purposely used an affine registration for three main reasons: (i) proving the robustness of the method also in case of roughly overlap between the anatomical regions of different TBI subjects; (ii) avoiding misregistration issues due to the particularly challenging process to apply a non linear transformation to a cohort with huge brain deformations; (iii) registering all the subjects to a common reference space keeping as much as possible the individual differences of the subjects and the relative lesions. Besides these initial steps, the analysis pipeline includes two principal sub-pipelines: a complex network pipeline and a machine learning pipeline that are schematized in **Figure 1** and are described in detail in the next two sections.

2.3. Multiplex Network Pipeline

After image processing, each scan was parceled in homologous non-overlapping parallelepipeds or patches of V voxels (where 1 voxel is 1 cubic millimeter) in order to obtain a 3D grid. These patches represent nodes of a brain network, and the absolute values of Pearson's correlation between patch pairs were considered the links between the nodes. In other words, each network link is obtained by computing the correlation voxel-by-voxel between the T1 intensities of two patches. Therefore, for each scan, we obtained a weighted brain network using a patch-based segmentation. To remove links due to the noise, we neglected all the correlations lower than 0.3. This threshold has not been chosen arbitrarily but has been demonstrated in our previous works (Amoroso et al., 2018b,c) as a threshold that maximizes classification performance and is the best trade-off between minimizing noise and maintaining effective network information in this multiplex network methodology. This threshold choice is also confirmed by other works in literature, for example, Mukaka (2012) suggests that correlations lower than 0.3 are negligible in his guide about the appropriate use of correlation coefficients in medical research. To further avoid false positive links in the networks, we also excluded the patches with a non-brain number of voxels exceeding 10% of their volume. The idea behind this study is that seizure development in TBI patients may be related to injury severity which, as many study demonstrate, results in diffuse cerebral edemas, hemorrhages, contusions, and distortions of brain tissue localized in multiple brain regions both close to and distant from the lesion area. Patch-based approach is aimed to detect this alterations in terms of correlation variation between regions with and without tissue damage over the TBI cohort (Kurland et al., 2012). A patch-based approach is a beneficial trade-off between a voxel-based approach and an ROI-based approach and has already been found to be beneficial in the field of other neurodegenerative diseases (Suk et al., 2014). It has three main advantages: (i) it overcomes the problem of the "curse of dimensionality," (ii) it does not depend on segmentation accuracy, and (iii) it is robust to misregistration errors (Amoroso et al., 2018a). Therefore, the patch representation can be very useful for TBI patients for whom ROI segmentation and spatial registration are particularly challenging tasks due to the large and irregular brain deformations caused by TBI lesions. To give a sense of the injury severity and the related processing challenges that were faced for this cohort, in **Figure 2**, for some of the TBI subjects with the most severe imaging findings, axial and coronal planes of brain scans after the processing (brain extraction and registration) are represented alongside the template that the subjects' scans are registered to.

The size V of each patch was varied from 1,000 to 8,000 voxels in steps of 1,000 to investigate the most appropriate scales to study seizure development in TBI patients. At different scales, the patch number is not constant but is determined by the patch size. The grid's origin is fixed for all the scales because we start to segment each image from the medial sagittal plane which separates the two brain hemispheres in order to uniformly cover each hemisphere with an equal number of rectangular boxes. For each scale, we built a multiplex network $G = \{G_1, G_2, ..., G_i, ..., G_M\}$ that is a collection of single subjects' weighted networks $G_i = (N, E, W)$ sharing the same nodes N, while the set of links E and weights W change depending on the subject's brain networks connections or layer connections. In other words, in each multiplex network, the same number of nodes (patches that each scan is segmented into) can be connected in different ways depending on the specific correlation coefficient values that characterize the network connections of a certain layer. Then, given N, the number of network nodes, we obtained $8N$ features for each subject, $4N$ features of single layer and $4N$ features of multiplex networks. The single layer features used in this work are strength and inverse participation ratio, given by the Equations (1) and (2), and their conditional means over the nodes with the same degree k, thus having the same connection number. Conditional means, given by the Equations (4) and (5), can be useful to examine whether, on average, the weights of central nodes and less connected nodes are identically distributed.

$$s_i^{\alpha} = \sum_{j=1}^{N} w_{ij}^{\alpha} \tag{1}$$

[2]https://qc.loni.usc.edu

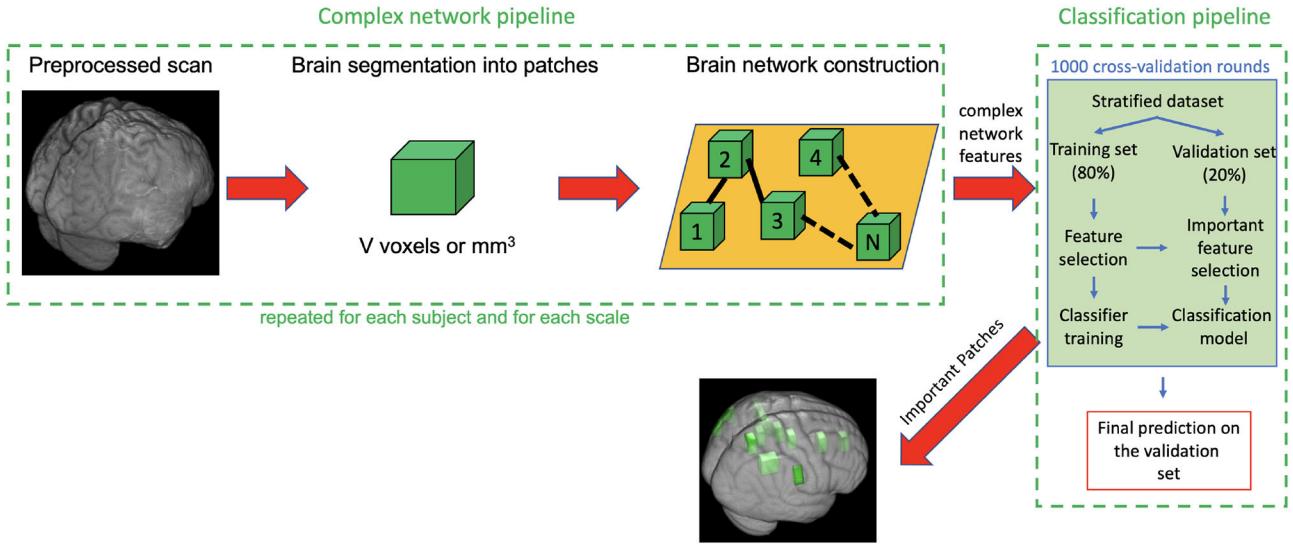


FIGURE 1 | Description of the complex network pipeline and machine learning pipeline. First each subject scan is preprocessed, segmented into patches, then for each subject a weighted undirected network was built and some complex network features were computed. Finally, the feature representation (subject × network features), obtained after the removal of null mean and variance features and highly correlated features, was used as input for the classification pipeline. The machine learning pipeline includes 1,000 rounds of cross-validation (CV). In each round the following steps are performed: (i) dataset was stratified; (ii) 80% of the stratified dataset was used as training set and 20% as validation set; (iii) training set was used, through a first nested Random Forest (RF) classifier, to select the most important features for the discrimination of seizure-free and seizure affected subjects: (iv) these selected features were used in turn to train a second RF classifier; (v) the important features and the classification models obtained on the training test were used to classify the subjects of the validation set; (vi) Averaging the classification performances over the 1,000 CV rounds, we obtained the final accuracy sensitivity, specificity, area under the receiver operating characteristics curve (AUC) and confidence interval on the validation test.

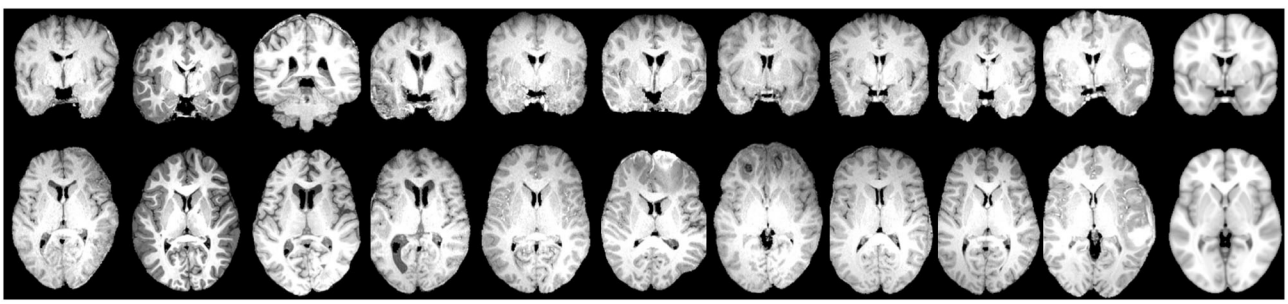


FIGURE 2 | Examples of coronal plane **(Top)** and axial plane **(Bottom)** of TBI subjects with the most severe imaging findings **(Left)** and the MNI152 reference template **(Right)** which the subjects' images are aligned to.

$$y_i^{\alpha} = \sum_{j=1}^{N} \left(\frac{w_{ij}^{\alpha}}{s_i^{\alpha}} \right)^2 \tag{2}$$

$$s(k)^{\alpha} = \frac{1}{N_k} \sum_{i=1}^{N} s_i^{\alpha} \delta(k_i^{\alpha}, k) \tag{3}$$

$$Y(k)^{\alpha} = \frac{1}{N_k} \sum_{i=1}^{N} Y_i^{\alpha} \delta(k_i^{\alpha}, k) \tag{4}$$

$\alpha = 1, .., M$ indicates the network layer, $w_i j$ is the correlation between gray level intensities of the nodes $i, j = 1, .., N$ with M subject number, N is the node number, N_k is the number of nodes having degree k, and δ is the Kronecker Delta function. Strength and inverse participation ratio indicate, respectively, the importance of a node and how evenly distributed the connections between nodes are. Specifically, $(y_i^{\alpha})^{-1} \in (1, k_i^{\alpha})$ has value k_i^{α} if the weights of the links of node i are distributed uniformly and, it has value 1 if the weight of one link is much larger than the other weights (Bianconi, 2018). The multiplex network features were obtained by weighing the previous quantities on the multiplex network degree k_{multi}, given

by (5), indicating the number of connections of a node in the multiplex network.

$$k_i^{multi} = \sum_{j=1}^{N} a_{ij}^{multi}; \quad (5)$$

where a_{ij}^{multi} is 1 if there is a layer with at least one link between nodes i and j and zero otherwise as described in Amoroso et al. (2018b). From now on, we will refer to these multiplex quantities as multi-strength, multi-inverse participation ratio, conditional strength, and conditional multi-inverse participation ratio. Overall, we obtained for each scale V, a M (subject network number) X 8N feature representation to analyze with the machine learning pipeline described in the next section.

2.4. Machine Learning Pipeline

For each scale V, multiplex network features were used to train a Random Forest (RF) classifier and obtain reliable classification models to identify, on the validation set, which TBI patients have developed seizures and which have not. This classification process, preceded by the removal of null mean and variance features, and highly correlated features (> 0.95), was carried out within a machine learning pipeline that includes 1,000 rounds of stratified cross-validation (Vabalas et al., 2019). For each round, we randomly picked the same percentage of seizure-free subjects and seizure-affected subjects in order to examine balanced datasets. After the stratification, a training set (80% of the stratified set) and validation set (20% of the stratified set) were defined. Subsequently, we first used a nested RF classifier, on the training set, to select and record features exceeding the third quartile of the importance distribution computed in terms of mean accuracy decrease. Then, we used those features to train a second RF classifier and obtain the classification models. Feature selection and training phases were nested within each cross-validation round and were blind to the validation set to avoid the "double dipping" problem (Kriegeskorte et al., 2009). Finally, we used the classification models and the important features retrieved during the training phase to classify the two clinical classes. Classification performances for each scale were evaluated in terms of accuracy, specificity, sensitivity, and Area Under the receiver-operating-characteristic Curve (AUC) averaged over all the cross-validation rounds. For the average accuracy, we also reported the 95% confidence interval computed according to the Wilson score interval (Wilson, 1927). We chose to use RF model because it is a robust and easy-to-tune model, it does not overfit thanks to internal bagging and it is particularly appropriate for analyses with high-dimensional feature spaces and small sample sizes (even $<$ 100) (Biau and Scornet, 2016; Floares et al., 2017). Each forest was grown with 500 trees, a number large enough for the out-of-bag error to reach the typical training plateau (Breiman, 1996). Therefore, in the internal bagging, given the training set, 500 bootstraps are formed obtaining 500 new subjects sets used to grow 500 trees. Each tree is grown by randomly choosing a subset of features equal to the square root of the feature number. The learning model built in this way can be then used to compute the out-of-bag error and the accuracy on the data left out of the training set.

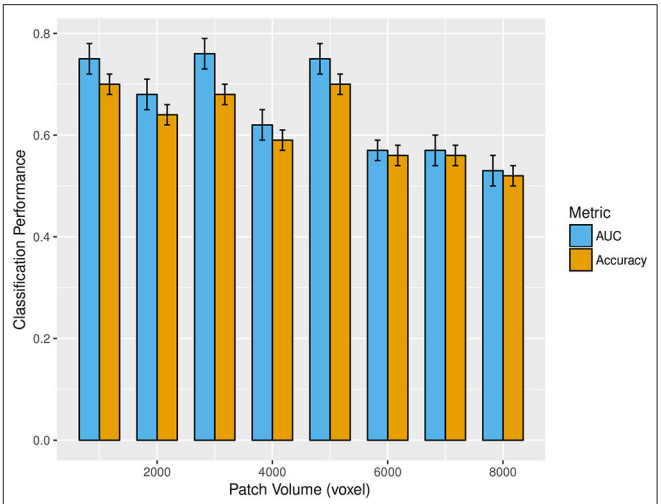


FIGURE 3 | The bar plot shows the mean and standard deviations of accuracy (gold) and AUC (cyan) over 1,000 rounds of cross-validation as patch volume changes (from 1,000 to 8,000 voxels). The best classification performances were obtained at 1,000, 3,000, and 5,000 voxels.

2.5. Important Feature Assessment

After having found for each round the important features, we evaluated, for the best scales, which feature occurrences had not happened by chance over the 1,000 rounds by using the statistical test of equal or given proportions (Newcombe, 1998). Therefore, the most important features over all the cross-validation rounds were found by considering, after the Bonferroni correction for multiple comparison, a $p - value < \alpha * N^{-1}$ with $\alpha = 0.05$. From the most important nodal features, it was possible to find the most important network nodes or patches, and thus the most important anatomical regions. We considered an anatomical region significantly related to the seizure development only if it occupied an important patch with a volume greater than 10% of the patch voxels. To identify the most important anatomical regions, we used Talairach labels projected in MNI 152 space (Lancaster et al., 2000). Other details on the reliability of the feature selection methods used in this work are discussed in the **Supplementary Material**.

3. RESULTS

3.1. Classification Performance and Feature Evaluation

Figure 3 shows the mean and standard deviations of accuracy and AUC over all the cross-validation rounds as a function of the patch size. The best classification performances were found at three patch volumes: 1, 000, 3, 000, and 5, 000 voxels.

Accuracy, specificity, sensitivity, and AUC with the corresponding standard deviations obtained for these three optimal scales are reported in **Table 3**.

Even though we found the best accuracy for a patch size of 1,000 and 5,000 voxels (with a 95% confidence interval of [67, 73%]), and the best AUC for 3,000 voxels, the classification performances at these three scales are statistically comparable. Another important aspect is to examine which network properties are more important to discriminate the two clinical groups. In this regard, we evaluated the mean percentage of features associated with a certain network metric that are selected as important in a cross validation-round. In **Figure 4**, the mean percentage of features selected over the cross-validation rounds and relative to each of the eight network metrics, is reported for the three most informative scales. This experiment was performed without excluding from the classification the highly correlated features.

Even though, for each scale, all metrics extracted contribute to the discrimination of the two clinical groups, we can notice that the nodal metrics have a greater relevance compared with the conditional quantities.

TABLE 3 | Accuracy, specificity, sensitivity, and AUC with the relative standard deviations obtained at the scales of 1,000, 3,000, and 5,000 voxels to which the best classification performances were reached.

Patch volume	Accuracy	Specificity	Sensitivity	AUC
1,000 voxels	**0.70 ± 0.03**	**0.74 ± 0.04**	0.66 ± 0.04	0.75 ± 0.02
3,000 voxels	0.68 ± 0.03	0.70 ± 0.04	0.67 ± 0.04	**0.76 ± 0.02**
5,000 voxels	**0.70 ± 0.03**	0.68 ± 0.04	**0.69 ± 0.04**	0.75 ± 0.02

The highest value for the four classification metrics are reported in bold.

3.2. ROI vs. Patch-Based Network Approach

We also compared the patch-based network approach with a standard ROI-based approach to evaluate the efficacy of the proposed complex network methodology to predict seizure development in TBI patients. We used the publicly available brain segmentation package, FreeSurfer (FS) v.6.0 (Fischl, 2012), which automatically performs: brain extraction, intensity normalization, spatial registration, volume labeling, segmentation, and all steps necessary to compute morphological features from each image. This tool allowed us to obtain 182 features, for each MRI scan, including subcortical and cortical gray matter parcellations, white matter parcellations, total gray and white matter volumes, and intracranial volume. These FS features were then used to distinguish TBI subjects who have developed epilepsy from those who have not by adopting the same machine learning pipeline used for the complex

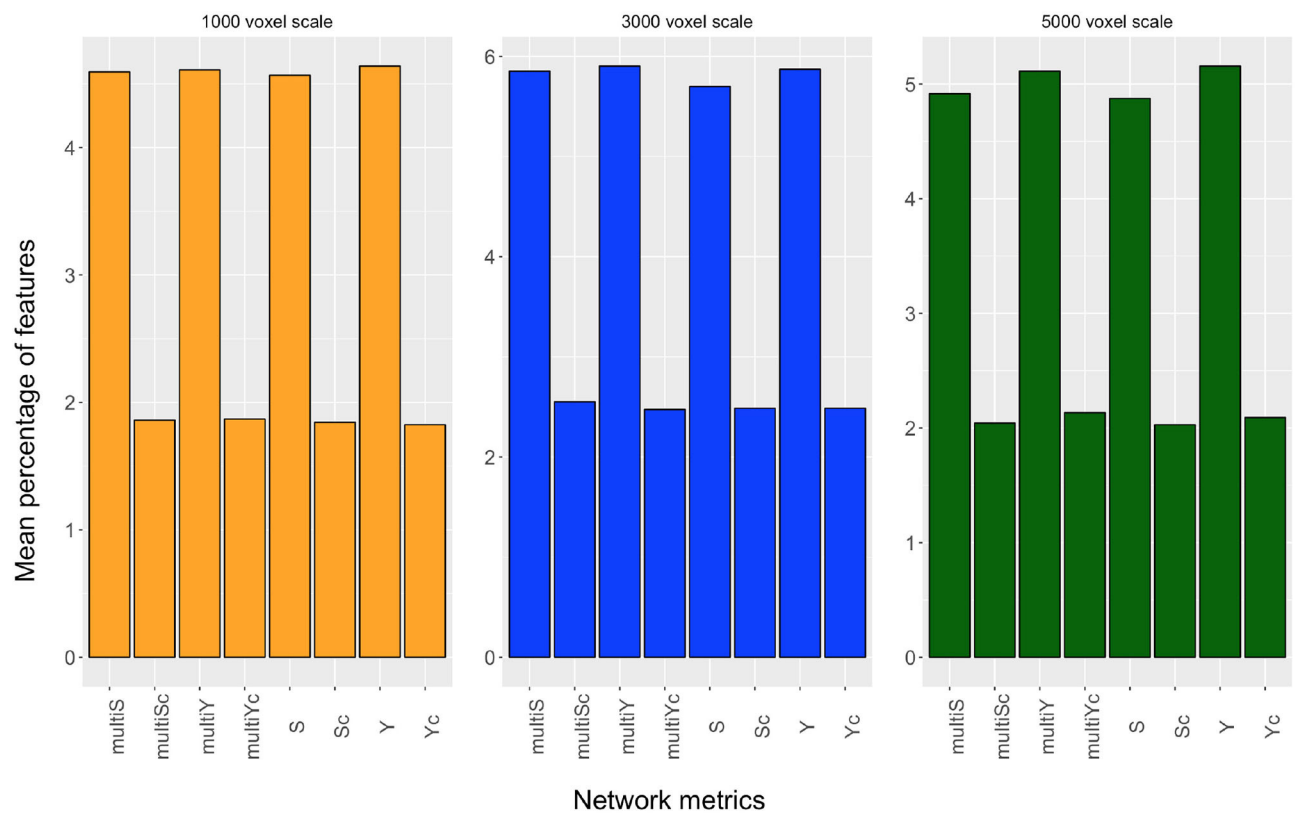


FIGURE 4 | Mean percentage of features, selected over the cross-validation rounds, relative to strength (S), inverse participation ratio (Y), multi-strength (multiS), inverse-participation ratio (multiY), and their conditional means(Sc, Yc, multiSc, multiYc) are reported for the scales of 1,000 voxels (yellow barplot), 3,000 voxels (blue barplot), and 5,000 voxels (dark green barplot).

network features. In **Figure 5**, receiver operating characteristics (ROC) curve and the related area under the curve (AUC) are reported for the three best scales of the complex networks and for FS.

We can notice that network approach outperform FS approach (accuracy: 0.66±0.02, sensitivity: 0.69±0.04, specificity: 0.62 ± 0.04).

3.3. Anatomical Regions Related to Seizure Development

For the three scales proved to be more appropriate to identify brain network alterations related to seizure development, we reported, in **Figure 6**, the brain areas (highlighted in green) corresponding to the most significant complex network features for the classification of seizure-free subjects and subjects with one seizure.

At the scale of 1,000 voxels, the significant patches ($p < 1.563 * 10^{-5}$ after the Bonferroni correction) identify anatomical regions mostly located in the right and left superior temporal gyrus lobe, but there are significant patches also in the left middle temporal gyrus, left inferior frontal and precentral gyrus, and in the right cerebellum within posterior lobe. At the scale of 3,000, the most important brain area ($p < 3.962 * 10^{-5}$ after the Bonferroni correction) for the two group discrimination corresponds to the cingulate gyrus in the left parietal lobe and in the right and left limbic lobe, sub-gyral in left and right frontal and parietal lobe, right and left precuneus, right postcentral gyrus, left inferior parietal lobule, angular gyrus, medial frontal gyrus, and superior occipital gyrus. Finally, the important areas ($p < 6.361 * 10^{-5}$ after the Bonferroni correction) at the scale of 5,000 voxels were the left and right cerebellum in the posterior lobe, right parahippocampal gyrus, right subcallosal gyrus, left inferior middle, and superior frontal gyrus, sub-gyral in the left frontal lobe, cingulate gyrus in the left limbic lobe, right and left extra-nuclear white matter, and left insula.

To make more understandable the relationship between patches and complex network features, in **Figure 7**, the distribution of the reciprocal of the inverse participation ratio of a patch, located in the left frontal lobe, is reported for both clinical groups. In the same figure, the representation of such a patch in a seizure-affected patient who has significant abnormalities in that area and in a seizure-free patient who does not have visible anomalies in that area is shown. The inverse participation ratio relative to the patch represented in **Figure 7** is an example of a network feature which is important for the discrimination of the two clinical groups. Indeed, from the box plot, we can notice that the median of the distribution for the seizure-affected subjects is significantly greater than the median of the distribution for the seizure-free subjects.

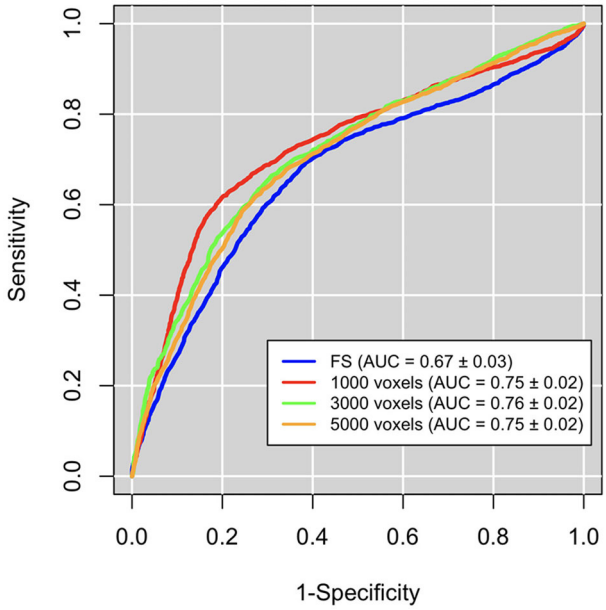


FIGURE 5 | Classification performances in terms of area under the receiver operating characteristics curve (AUC). Performaces obtained with the complex network features at the the best three scales (red, orange, and green curves) are significantly greater than those obtained with FreeSurfer (FS) features (blue curves).

4. DISCUSSION

In this work, an innovative multiplex network approach was used to find informative complex network features that can be used from machine learning systems for the identification of patients who have developed a seizure after a TBI. To the best of our knowledge, this is the first study to distinguish seizure-free subjects and seizure-affected subjects with an accuracy of 70% and an AUC of 76% by using T1-weighted MRI data. In Messori et al. (2005), PTE prediction using human MRI is based only on statistical evaluations. Classification performances obtained in this work are comparable and even higher than those found in La Rocca et al. (2019) and Garner et al. (2019) that examine functional and structural alterations related to seizure onset. All the network properties are proved to be useful to the classification, however, features relative to conditional metrics were selected less frequently in each cross-validation round. Classification performances were computed as a function of the patch volume, which was varied in an intermediate range (from 1,000 to 8,000 voxels) in order to avoid that the analysis was affected by a low sensitivity to subtle pathological changes in case of too large patches or by the 'curse of dimensionality' and misregistration in case of too small patches. The best classification performances were obtained at three different scales: 1,000, 3,000, and 5,000 voxels, proving that the study of seizure development in TBI patients requires multivariate analyses. Although some of the important anatomical regions, such as cingulate gyrus, sub-gyral, inferior central gyrus, and cerebellar tonsil in the right cerebellum are accordant for the three scales, we can notice that some morphological changes between the two clinical groups can be detected only at specific scales. This suggests that seizure development in TBI patients cannot be studied considering a unique scale, which is reasonable given the heterogeneity of the epileptogenesis process after a

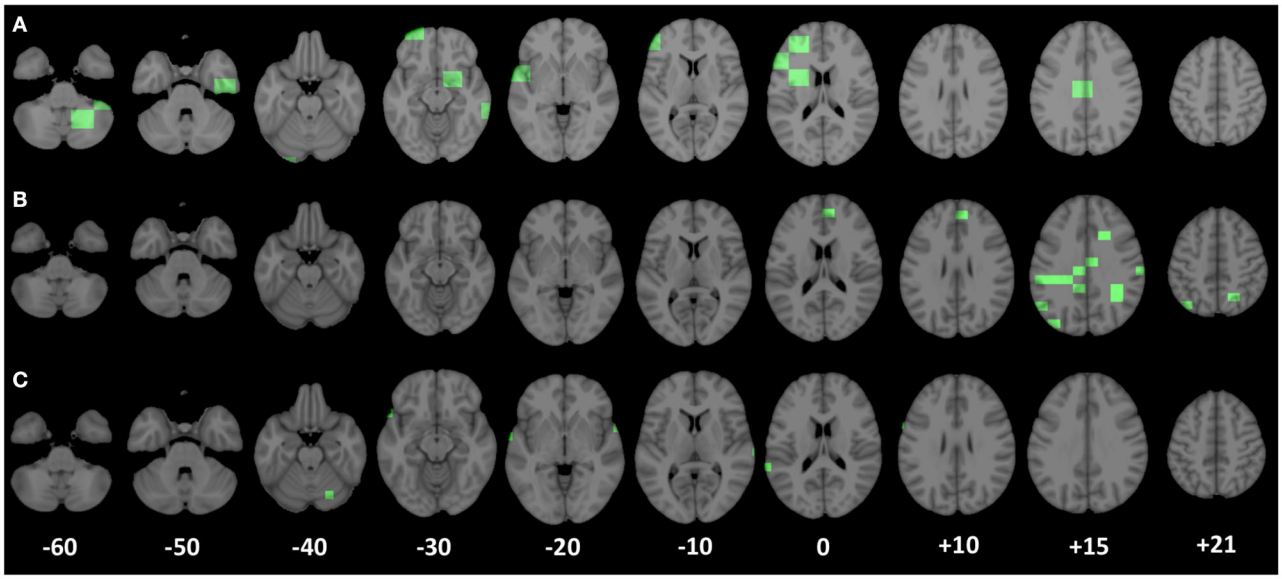


FIGURE 6 | Patches corresponding to the most important nodal complex network features for the discrimination of the two clinical groups (seizure-free and seizure-affected) are underlined in green along the axial planes of the MNI 152 template. **(A–C)** Display the most significant patches relative to the scales of 5,000, 3,000, and 1,000 voxels, respectively.

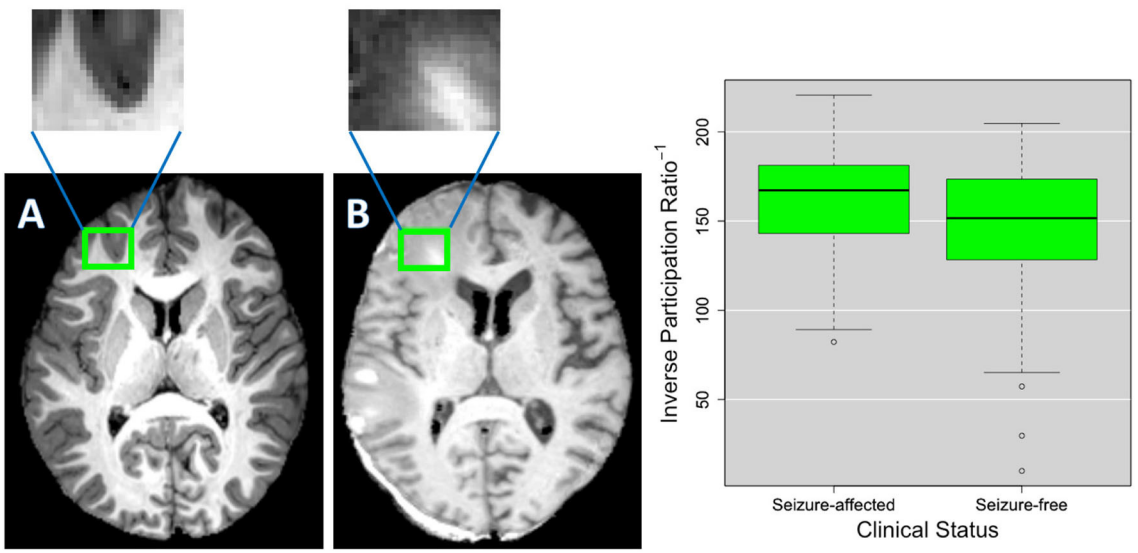


FIGURE 7 | As an example, **(A,B)** show the details of a patch in two different TBI subjects. As shown in the green box plots on the right, this patch has an inverse participation ratio that is significantly different ($p < 2.2 * 10^{-16}$) in the two clinical groups (seizure-free and seizure-affected patients). **(B)** Shows the patch pinpointing an area where voxels intensity is altered by a lesion and the surrounding edema and **(A)** shows the patch covering a brain area that does not have evident alterations.

TBI and the fact that TBIs affect the brain in different areas and at different scales. As a consequence, analyzing multiple scales can give a more exhaustive detection of the MRI changes related to seizure development after a TBI that a unique scale is not able to provide. Therefore, our methodology can be very useful to perform a multivariate analysis that take into account multi-scale features. Conventional volumetric analyses are based on manual segmentation of the Region-Of-Interest (ROI) that is time-consuming and affected by personal bias. The modern automated algorithms that allow the determination of volumes, thickness, and shape of anatomical structures often fail because of large lesion size and extensive tissue damage in TBI patients. In this regard, we demonstrated that our approach (AUC of 76%) is more effectiveness than a ROI-based approach like FS (AUC

of 67%). EpiBioS4Rx is an ongoing study that will enroll 300 patients, therefore in upcoming years, with a larger and more representative training sample, we will be able to fully exploit machine learning potentialities and obtain conclusive results about the generalization power of the model in predicting seizure development in TBI patients (Figueroa et al., 2012). Once more subjects will be enrolled and longitudinally examined, it will be also interesting to see if the proposed methodology is able to distinguish among immediate, early, and late seizures in order to take into account also the temporal aspect of the epileptogenic process. Besides, the completely automated complex network approach used in this work can be really beneficial because it allows an unsupervised identification of the important brain areas without being affected by ROI segmentation mistakes and time-consuming procedures. This methodology offers also other two main advantages: (i) it allows the identification of MRI changes which differentiate seizure-free and seizure-affected patients and which cannot be underlined using only CT MRI findings (see **Table 1**); (ii) it allows the identification of the brain scales at which the pathological changes related to seizure development occur. Indeed, as might be expected, given TBI variability, epileptogenesis mechanism will depend on alterations that happen at different scales (Cloots et al., 2013). It is interesting to notice that at the scale of 1,000 voxels, most of the patches are located at the periphery of the brain. This may be due to brain surface deformations or to subdural and epidural hematomas that are reported among the risk factors to develop epilepsy and are present in many subjects of this cohort as reported in **Table 1** (Agrawal et al., 2006). An example of subdural hemotoma is shown in **Figure 7B** in the left posterior part of the brain. Most of the clinical results are in line with recent studies about seizure development. Norden and Blumenfeld (2002) states the increased likelihood of cerebellar alterations in patients with epilepsy. In Shultz et al. (2013), MRI alterations were found in hippocampus subfields of rodents with epilepsy after a lateral fluid percussion injury. Tubi et al. (2019) showed that subjects with lesions in the temporal lobe are at high risk to develop epilepsy, suggesting that morphological alterations in the temporal lobe may play a strategic role in seizure occurrence. Hippocampus, cingulate gyrus, precentral gyrus, postcentral gyrus, and middle and inferior frontal gyrus were proved to be regions related to the epileptogenesis process also in studies that apply machine learning techniques to fMRI and sMRI to characterize patients with epilepsy (Zhang et al., 2012; Garner et al., 2019; La Rocca et al., 2019). It is worthwhile to notice that the highly correlated (> 0.95) features that were excluded in the classification correspond to complex network metrics related to the same patch and thus, to the same brain area. This ensures that we did not exclude any important region in the clinical validation and suggests that for some patches more complex network metrics are accordant with each other.

5. CONCLUSION

We have demonstrated that the combined use of complex networks and machine learning techniques can be useful to study seizure development in TBI patients. Multiplex networks were able to provide network features that allow us to distinguish TBI patients who have developed epilepsy from those who have not with an accuracy of 70%. In addition, a patch-based approach used to build the multiplex networks made it possible to identify, in an unsupervised way, the brain areas important for the discrimination of the two clinical groups, even though a perfect solution of optimum features is a challenging and still open matter. EpiBioS4Rx is an ongoing study that will enroll 300 patients, thus in the near future, a larger dataset will be available, and we will be able to obtain more conclusive results.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: access to data must be requested and approved by the EpiBioS4Rx steering committee. Requests to access these datasets should be directed to epibiossteeringcommittee@loni.usc.edu.

ETHICS STATEMENT

This work was approved by the UCLA Institutional Review Board (IRB# 16-001 576) and the local review boards at each EpiBioS4Rx Study Group institution. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ML conceived and conducted the analyses. ML, RG, NA, EL, MM, PV, AT, and DD analyzed the results and reviewed the manuscript.

FUNDING

This study was conducted with the support of the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH) under award numbers U54 NS100064 (EpiBioS4Rx) and R01NS111744.

ACKNOWLEDGMENTS

Data used in the preparation of this article were obtained from the The Epilepsy Bioinformatics Study for Antiepileptogenic Therapy (EpiBioS4Rx) database (https://epibios.loni.usc.edu). EpiBioS4Rx was funded by the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH) in 2017. EpiBioS4Rx is a large, international, multi-site Center without Walls (CWOW) which has been collecting longitudinal EEG, imaging, and blood data from human patients and an animal model with the primary goal to identify biomarkers of epileptogenesis after a traumatic brain injury and then provide therapies and treatments that may stop the development of post-traumatic epilepsy. The Principal Investigators of this initiative are Jerome Engel Jr.,

MD, Ph.D., Ronald Reagan University of California Los Angeles Medical Center; Aristea Galanopoulou, Ph.D., Albert Einstein College of Medicine; Solomon L. Moshé, MD, Albert Einstein College of Medicine; Terence O'Brien, MD, The University of Melbourne; Asla Pitkänen, Ph.D., University of Eastern Finland, Kuopio; AT Ph.D., University of Southern California; Paul M. Vespa, MD, Ronald Regan University of California Los Angeles Medical Center. EpiBioS4Rx is the result of efforts of many investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 30 sites across the world. EpiBioS4Rx data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California. A complete listing of EpiBioS4Rx investigators can be found at: https://epibios.loni.usc.edu.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.591662/full#supplementary-material

REFERENCES

Agrawal, A., Timothy, J., Pandit, L., and Manju, M. (2006). Post-traumatic epilepsy: an overview. *Clin. Neurol. Neurosurg.* 108, 433–439. doi: 10.1016/j.clineuro.2005.09.001

Amoroso, N., La Rocca, M., Bellotti, R., Fanizzi, A., Monaco, A., Tangaro, S., et al. (2018a). Alzheimer's disease diagnosis based on the hippocampal unified multi-atlas network (human) algorithm. *Biomed. Eng. Online* 17:6. doi: 10.1186/s12938-018-0439-y

Amoroso, N., La Rocca, M., Bruno, S., Maggipinto, T., Monaco, A., Bellotti, R., et al. (2018b). Multiplex networks for early diagnosis of Alzheimer's disease. *Front. Aging Neurosci.* 10:365. doi: 10.3389/fnagi.2018.00365

Amoroso, N., La Rocca, M., Monaco, A., Bellotti, R., and Tangaro, S. (2018c). Complex networks reveal early MRI markers of Parkinson's disease. *Med. Image Anal.* 48, 12–24. doi: 10.1016/j.media.2018.05.004

Bharath, R. D., Panda, R., Raj, J., Bhardwaj, S., Sinha, S., Chaitanya, G., et al. (2019). Machine learning identifies "rsfMRI epilepsy networks" in temporal lobe epilepsy. *Eur. Radiol.* 29, 1–10. doi: 10.1007/s00330-019-5997-2

Bianconi, G. (2018). *Multilayer Networks: Structure and Function*. Oxford: Oxford University Press. doi: 10.1093/oso/9780198753919.001.0001

Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test* 25, 197–227. doi: 10.1007/s11749-016-0481-7

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Cloots, R. J., Van Dommelen, J., Kleiven, S., and Geers, M. (2013). Multi-scale mechanics of traumatic brain injury: predicting axonal strains from head loads. *Biomech. Model. Mechanobiol.* 12, 137–150. doi: 10.1007/s10237-012-0387-6

Diaz-Arrastia, R., Agostini, M. A., Madden, C. J., and Van Ness, P. C. (2009). Posttraumatic epilepsy: the endophenotypes of a human model of epileptogenesis. *Epilepsia* 50, 14–20. doi: 10.1111/j.1528-1167.2008.02006.x

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Med. Inform. Decis. Mak.* 12:8. doi: 10.1186/1472-6947-12-8

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Floares, A., Ferisgan, M., Onita, D., Ciuparu, A., Calin, G., and Manolache, F. (2017). The smallest sample size for the desired diagnosis accuracy. *Int. J. Oncol. Cancer Ther.* 2, 13–19.

Garner, R., La Rocca, M., Barisano, G., Toga, A. W., Duncan, D., and Vespa, P. (2019). "A machine learning model to predict seizure susceptibility from resting-state fMRI connectivity," in *Proceedings of the Modeling and Simulation in Medicine Symposium*, 14 (Tucson, AZ: Society for Computer Simulation International). doi: 10.23919/SpringSim.2019.8732859

Humphreys, I., Wood, R. L., Phillips, C. J., and Macey, S. (2013). The costs of traumatic brain injury: a literature review. *Clin. Econ. Outcomes Res.* 5:281. doi: 10.2147/CEOR.S44625

Immonen, R., Harris, N. G., Wright, D., Johnston, L., Manninen, E., Smith, G., et al. (2018). Imaging biomarkers of epileptogenecity after traumatic brain injury-preclinical frontiers. *Neurobiol. Dis.* 123, 75–85. doi: 10.1016/j.nbd.2018.10.008

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Kim, J. A., Boyle, E. J., Wu, A. C., Cole, A. J., Staley, K. J., Zafar, S., et al. (2018). Epileptiform activity in traumatic brain injury predicts post-traumatic epilepsy. *Ann. Neurol.* 83, 858–862. doi: 10.1002/ana.25211

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12:535. doi: 10.1038/nn.2303

Kurland, D., Hong, C., Aarabi, B., Gerzanich, V., and Simard, J. M. (2012). Hemorrhagic progression of a contusion after traumatic brain injury: a review. *J. Neurotrauma* 29, 19–31. doi: 10.1089/neu.2011.2122

La Rocca, M., Amoroso, N., Monaco, A., Bellotti, R., Tangaro, S., Initiative, A. D. N., et al. (2018). A novel approach to brain connectivity reveals early structural changes in Alzheimer's disease. *Physiol. Measure.* 39:074005. doi: 10.1088/1361-6579/aacf1f

La Rocca, M., Garner, R., Jann, K., Kim, H., Vespa, P., Toga, A. W., et al. (2019). "Machine learning of multimodal MRI to predict the development of epileptic seizures after traumatic brain injury," in *Medical Imaging with Deep Learning (MIDL) Abstract* (London).

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., et al. (2000). Automated talairach atlas labels for functional brain mapping. *Hum. Brain Mapp.* 10, 120–131. doi: 10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8

Lutkenhoff, E. S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J. D., Owen, A. M., et al. (2014). Optimized brain extraction for pathological brains (optibet). *PLoS ONE* 9:e115551. doi: 10.1371/journal.pone.0115551

Lutkenhoff, E. S., Shrestha, V., Tejeda, J. R., Real, C., McArthur, D. L., Duncan, D., et al. (2020). Early brain biomarkers of post-traumatic seizures: initial report of the multicentre epilepsy bioinformatics study for antiepileptogenic therapy (epibios4rx) prospective study. *J. Neurol. Neurosurg. Psychiatry.* 91, 1154–1157. doi: 10.1136/jnnp-2020-322780

Messori, A., Polonara, G., Carle, F., Gesuita, R., and Salvolini, U. (2005). Predicting posttraumatic epilepsy with MRI: prospective longitudinal morphologic study in adults. *Epilepsia* 46, 1472–1481. doi: 10.1111/j.1528-1167.2005.34004.x

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24, 69–71. doi: 10.4314/mmj.v20i1.10949

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat. Med.* 17, 873–890. doi: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Norden, A. D., and Blumenfeld, H. (2002). The role of subcortical structures in human epilepsy. *Epilep. Behav.* 3, 219–231. doi: 10.1016/S1525-5050(02)00029-X

Shah, P., Bassett, D. S., Wisse, L. E., Detre, J. A., Stein, J. M., Yushkevich, P. A., et al. (2019). Structural and functional asymmetry of medial temporal subregions in unilateral temporal lobe epilepsy: a 7T MRI study. *Hum. Brain Mapp.* 40, 2390–2398. doi: 10.1002/hbm.24530

Shultz, S. R., Cardamone, L., Liu, Y. R., Hogan, R. E., Maccotta, L., Wright, D. K., et al. (2013). Can structural or functional changes following traumatic brain injury in the rat predict epileptic outcome? *Epilepsia* 54, 1240–1250. doi: 10.1111/epi.12223

Suk, H.-I., Lee, S.-W., Shen, D., and Alzheimers Disease Neuroimaging Initiative. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* 101, 569–582. doi: 10.1016/j.neuroimage.2014.06.077

Tubi, M. A., Lutkenhoff, E., Blanco, M. B., McArthur, D., Villablanca, P., Ellingson, B., et al. (2019). Early seizures and temporal lobe trauma predict post-traumatic epilepsy: a longitudinal study. *Neurobiol. Dis.* 123, 115–121. doi: 10.1016/j.nbd.2018.05.014

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14:e0224365. doi: 10.1371/journal.pone.0224365

Verellen, R. M., and Cavazos, J. E. (2010). Post-traumatic epilepsy: an overview. *Therapy* 7:527. doi: 10.2217/thy.10.57

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* 22, 209–212. doi: 10.1080/01621459.1927.10502953

Zhang, J., Cheng, W., Wang, Z., Zhang, Z., Lu, W., Lu, G., et al. (2012). Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy. *PLoS ONE* 7:e36733. doi: 10.1371/journal.pone.0036733

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 22 December 2020
doi: 10.3389/fnins.2020.610239

RatLesNetv2: A Fully Convolutional Network for Rodent Brain Lesion Segmentation

Juan Miguel Valverde[1]*, *Artem Shatillo*[2], *Riccardo De Feo*[1,3,4], *Olli Gröhn*[1], *Alejandra Sierra*[1] and *Jussi Tohka*[1]

[1] A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland, [2] Charles River Discovery Services, Kuopio, Finland, [3] Centro Fermi-Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi, Rome, Italy, [4] Sapienza Università di Roma, Rome, Italy

We present a fully convolutional neural network (ConvNet), named RatLesNetv2, for segmenting lesions in rodent magnetic resonance (MR) brain images. RatLesNetv2 architecture resembles an autoencoder and it incorporates residual blocks that facilitate its optimization. RatLesNetv2 is trained end to end on three-dimensional images and it requires no preprocessing. We evaluated RatLesNetv2 on an exceptionally large dataset composed of 916 T2-weighted rat brain MRI scans of 671 rats at nine different lesion stages that were used to study focal cerebral ischemia for drug development. In addition, we compared its performance with three other ConvNets specifically designed for medical image segmentation. RatLesNetv2 obtained similar to higher Dice coefficient values than the other ConvNets and it produced much more realistic and compact segmentations with notably fewer holes and lower Hausdorff distance. The Dice scores of RatLesNetv2 segmentations also exceeded inter-rater agreement of manual segmentations. In conclusion, RatLesNetv2 could be used for automated lesion segmentation, reducing human workload and improving reproducibility. RatLesNetv2 is publicly available at https://github.com/jmlipman/RatLesNetv2.

Keywords: ischemic stroke, lesion segmentation, deep learning, rat brain, magnetic resonance imaging

OPEN ACCESS

Edited by:
Tim B. Dyrby,
Technical University of Denmark, Denmark

Reviewed by:
Yi Zhang,
Zhejiang University, China
Shanshan Jiang,
Johns Hopkins Medicine, United States

***Correspondence:**
Juan Miguel Valverde
juanmiguel.valverde@uef.fi

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *25 September 2020*
Accepted: *25 November 2020*
Published: *22 December 2020*

Citation:
Valverde JM, Shatillo A, De Feo R, Gröhn O, Sierra A and Tohka J (2020) RatLesNetv2: A Fully Convolutional Network for Rodent Brain Lesion Segmentation. Front. Neurosci. 14:610239. doi: 10.3389/fnins.2020.610239

1. INTRODUCTION

Rodents frequently serve as models for human brain diseases. They account for more than 80% of the animals used in research in recent years (Dutta and Sengupta, 2016). In addition to basic research, rodent models are important in, for example, drug discovery and the development of new treatments. *In vivo* imaging of rodents is used for monitoring disease progression and therapeutic response in longitudinal studies. In particular, magnetic resonance imaging (MRI) is essential in pre-clinical studies for conducting quantitative analyses due to its non-invasiveness and versatility. As an example, the quantification of brain lesions requires segmenting the lesions, and the lack of reliable tools to automate rodent brain lesion segmentation forces researchers to segment these images manually.

Manual segmentation can be prohibitively time-consuming as studies involving animals may acquire hundreds of three-dimensional (3D) images. Furthermore, the difficulty of defining lesion boundaries leads to moderate inter- and intra-rater agreement; previous studies have reported that Dice coefficients (Dice, 1945) between annotations made by two humans can be as low as 0.73

(Valverde et al., 2019) or 0.79 (Mulder et al., 2017a). Moderate inter-rater agreement is caused by several factors that affect the segmentation quality, including partial volume effect, image contrast and annotator's knowledge and experience. Despite these liabilities, manual segmentation is the gold standard and a common practice among researchers who use animal models (Moraga et al., 2016; De Feo and Giove, 2019).

Semi-automatic methods are a faster alternative to manual segmentation. However, they fail to overcome the subjectivity of the manual segmentation, as human interaction is required. To the best of the authors' knowledge, there are only two studies that introduce and evaluate semi-automatic algorithms for rodent brain lesion segmentation. Wang et al. (2007) evaluated a combination of thresholding operations commonly used in the literature to segment lesions on apparent diffusion coefficient (ADC) maps and T2-weighted images. Choi et al. (2018) first normalized the intensity values of each image with respect to the contralateral hemisphere of the brain, and they performed a series of thresholding operations to segment permanent middle cerebral artery occlusion ischemic lesions in 31 diffusion-weighted images (DWIs) of the rat brain. Both methods require the manual segmentation of the contralateral hemisphere. Additionally, these thresholding-based and other voxel-wise approaches disregard the spatial and contextual information of the images, and they are sensitive to the image modality, contrast, and possible artifacts. Pipelines that rely on thresholding operations may result in poor and inconsistent segmentation results in the form of holes within and outside the lesion mask (**Figure 1**).

For lesion segmentation in rodent MRI, researchers have proposed a few fully-automated methods in recent years. Mulder et al. (2017a) developed a level-set-based algorithm that was tested on 121 T2-weighted mouse brain scans. However, the accuracy of their method heavily relies on the performance of other independent steps, such as registration, skull-stripping and contralateral ventricle segmentation. Arnaud et al. (2018) derived a pipeline that detects voxels that are anomalous with respect to a reference model of healthy animals, and they evaluated the pipeline on 53 rat brain MRI maps. Nonetheless, this pipeline was specifically designed for quantitative MRI, and it expects sham-operated animals in the data set, a requirement that is not always feasible.

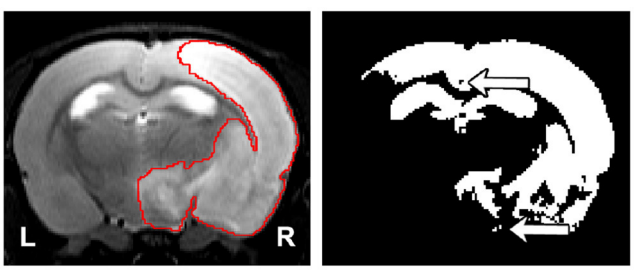


FIGURE 1 | (Left) Representative lesion with its ground truth. (Right) Segmentation of the lesion using thresholding where the threshold was found by maximizing the Dice coefficient with respect to the manual segmentation. The arrows indicate the presence of holes and islands (independently connected components) within and outside the mask, respectively. The hippocampus and ventricles were entirely misclassified as lesion.

Deep learning, and more specifically convolutional neural networks (ConvNets), has become increasingly popular due to its competitive performance in medical image segmentation. Literature on brain lesion segmentation in MR images with ConvNets is dominated by approaches tested on human-derived data (e.g., Duong et al., 2019; Gabr et al., 2019; Yang et al., 2019). Despite using ConvNets, typical brain lesion segmentation approaches are multi-step, i.e., they rely on preprocessing procedures, such as noise reduction, registration, skull-stripping and inhomogeneity correction. Therefore, the performance of the preprocessing steps influences the quality of the final segmentation. In contrast to human-derived data, rodent segmentation data sets are scarce and smaller in size (Mulder et al., 2017b); consequently, ConvNet-based segmentation methods benchmarked on rodent MR images are rare. An exception—not in the lesion segmentation—is Roy et al. (2018)'s work, which introduced a framework to extract brain tissue (i.e., skull-stripping) on human and mice MRI scans after traumatic brain injury.

We present RatLesNetv2, the first 3D ConvNet for segmenting rodent brain lesions in pre-clinical MR images. Our fully-automatic approach is trained end to end, requires no preprocessing, and it was validated on a large and diverse data set composed by 916 MRI rat brain scans at nine different lesion stages from 671 rats utilized to study focal cerebral ischemia. We extend our earlier conference paper (Valverde et al., 2019) by (1) improving our previous ConvNet (Valverde et al., 2019) with a deeper and different architecture and providing an ablation study (Meyes et al., 2019) justifying certain architectural choices; (2) evaluating the generalization capability of our model on a considerably larger and more heterogeneous data set via Dice coefficient, compactness and Hausdorff distance under different training settings (training set size and different ground truth); and (3) making RatLesNetv2 publicly available.

We show that RatLesNetv2 generates more realistic segmentations than our previous RatLesNet, and than 3D U-Net (Çiçek et al., 2016) and VoxResNet (Chen et al., 2018a), two state-of-the-art ConvNets specifically designed for medical image segmentation. Additionally, the Dice coefficients of the segmentations derived with RatLesNetv2 exceeded inter-rater agreement scores.

2. MATERIALS AND METHODS

2.1. Data

The data set consisted of 916 MR T2-weighted brain scans of 671 adult male Wistar rats weighting between 250 and 300 g. The data, provided by Discovery Services site of Charles River Laboratories,[1] were derived from 12 different studies. Transient (120 min) focal cerebral ischemia was produced by middle cerebral artery occlusion in the right hemisphere of the brain (Koizumi et al., 1986). MR data acquisitions were

[1]https://www.criver.com/products-services/discovery-services

TABLE 1 | Number of scans per study segregated by lesion stage, including sham-operated animals.

Study	2 h	24 h	D3	D7	D14	D21	D28	D35	Shams
A	12	12	0	0	0	0	0	0	24
B	0	46	0	0	0	0	0	0	3
C	0	59	0	0	0	0	0	0	1
D	0	162	0	0	0	0	0	0	4
E	0	0	0	0	0	0	0	20	1
F	0	33	30	0	30	0	27	0	46
G	0	0	0	53	0	0	0	0	12
H	0	45	0	0	0	0	0	0	0
I	0	0	64	0	0	0	62	0	0
J	0	32	0	0	0	0	0	0	0
K	0	17	0	0	0	0	0	0	0
L	0	0	41	0	0	40	0	0	40
Total	12	406	135	53	30	40	89	20	131

performed at different time-points after the occlusion (for details, see **Table 1**). Some studies also had sham-operated animals that underwent identical surgical procedures, but without the actual occlusion. All animal experiments were conducted according to the National Institute of Health (NIH) guidelines for the care and use of laboratory animals, and approved by the National Animal Experiment Board, Finland. Multi-slice multi-echo sequence was used with the following parameters; TR =2.5 s, 12 echo times (10–120 ms in 10 ms steps), and 4 averages in a horizontal 7T magnet. T2-weighted images were calculated as the sum of the all echoes. Eighteen coronal slices of 1 mm thickness were acquired using a field-of-view of 30 × 30 mm^2 producing 256 × 256 imaging matrices of resolution 117 × 117 μm. No MRI preprocessing steps, such as inhomogeneity correction, artifact removal, registration or skull stripping, were applied to the T2-weighted images. Images were zero-centered and their variance was normalized to one.

The provided lesion segmentations were annotated by several trained technicians employed by Charles River. We performed an additional independent manual segmentation of the lesions on the first study that was acquired (study A, **Table 1**) to approximate inter-rater variability. The average Dice coefficient (Dice, 1945) between the two manual segmentations was 0.67 with a standard deviation of 0.12 on 2 h lesions and 0.79 with a standard deviation of 0.08 on 24 h lesions. The overall average was 0.73 ± 0.12. Unless stated otherwise, we used our independent segmentation as the ground truth for study A.

We produced two different train/test set divisions. (1) In the first one, the training set contained the 48 scans of the study which was used to approximate inter-rater variability (study A, **Table 1**) and the test set contained the remaining 868 images. The training set was further divided to training (36 images) and validation sets (12 images). This train/test division is referred to as "**homogeneous**" and its train/validation split has the same ratio 2/24 h time-points and sham/no-sham animals. (2) The second division also contained 48 training scans and the test set contained 868 scans, but the training set was different from the homogeneous division. This division is referred to as "**heterogeneous**" because the training set was more diverse. The training set was divided into training (40 images) and validation (8 images) set. The training and the validation sets were formed by 5 and 1 images per lesion time-point, respectively, with no images from sham-operated animals. The size of our training set was deliberately much smaller than the test set for two reasons: (1) to replicate the typical pre-clinical setting in which rodent MR images are few and (2) to create a large and representative test set.

2.2. Convolutional Neural Networks

Convolutional neural networks (ConvNets) use stacks of convolutions to transform spatially correlated data, such as images, to extract their features. The first layers of the network capture low-level information, such as edges and corners, and the final layers extract more abstract features. The number of convolutions adjusts two attributes of ConvNets: parameter number and network depth. An excessive number of parameters leads to overfitting—memorizing the training data; an insufficient number of parameters constrains the learning capability of the model. Model depth is associated with the number of times the input data is transformed, and this depth also adjusts the area that influences the prediction—the receptive field (RF). Recent approaches reduce model parameters while maintaining the RF by using more stacked convolutions of smaller kernel size (Szegedy et al., 2016).

Model architectures based on U-Net (Ronneberger et al., 2015) are popular in medical image segmentation tasks. In contrast to patch-based models, the input images and the generated masks are the same size, which makes U-Nets computationally more efficient to train and to evaluate. The U-Net architecture resembles an autoencoder with skip connections between the same levels of the encoder and decoder. The encoder transforms and reduces the dimensionality of the input images, and the decoder recovers the spatial information with the help of skip connections.

Skip connections also facilitate the gradient flow during back-propagation (Drozdzal et al., 2016), but they are not sufficient to prevent the gradient of the loss to vanish, which makes the network harder to train. This is also referred as the vanishing gradient problem (He et al., 2016), and it particularly affects the final layers of the encoder part. Adding residual connections (He et al., 2016) along the network alleviates the vanishing gradient problem and it also yields in faster convergence rates during the optimization (Drozdzal et al., 2016).

2.3. RatLesNetv2 Architecture

RatLesNetv2 (**Figure 2**) has three downsampling and three upsampling stages connected via skip connections. Maxpooling downsamples the data with a window size and strides of 2, and trilinear interpolation upsamples the feature maps. Bottleneck layers (**Figure 2**, green blocks) stack a ReLU activation function, a batch normalization (BatchNorm) layer (Ioffe and Szegedy, 2015) and a 3D convolution with kernel size of 1 that combines and modifies the number of channels of the feature maps from *in* to *out*. ResNetBlock layers (**Figure 2**, orange blocks) contain two stacks of ReLU activations, BatchNorm, and 3D convolutions

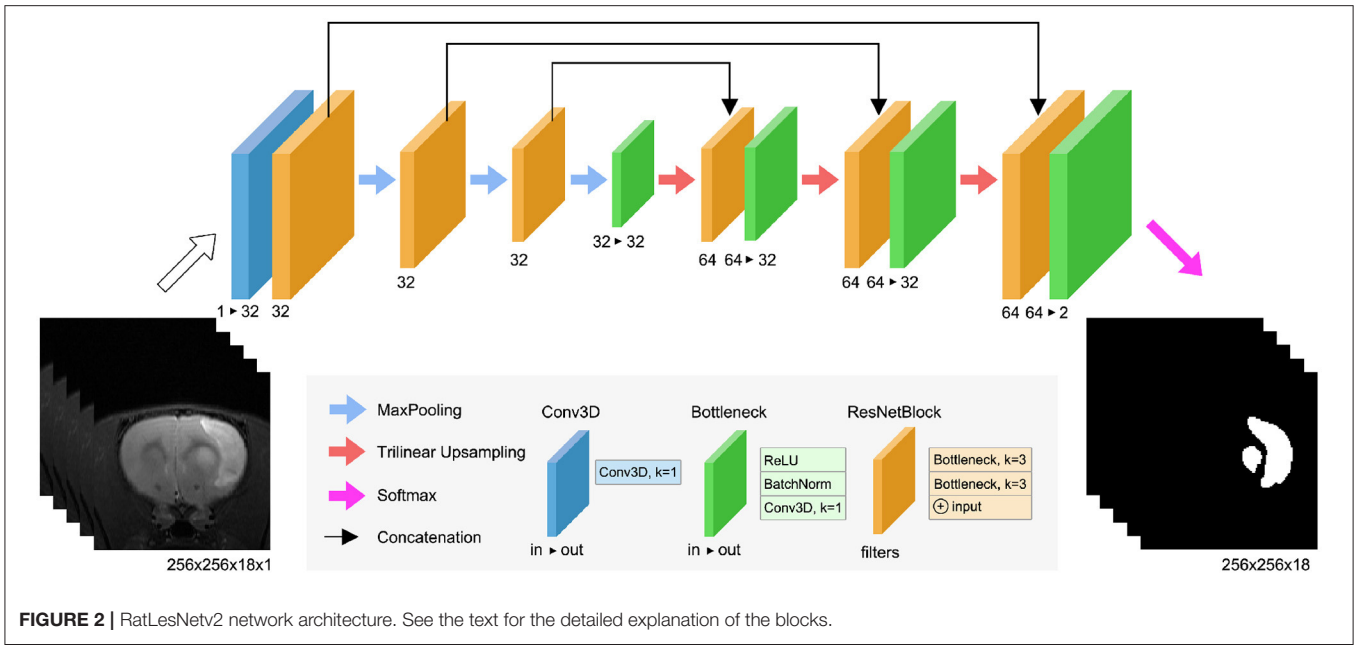


FIGURE 2 | RatLesNetv2 network architecture. See the text for the detailed explanation of the blocks.

with kernel size of 3. Similarly to VoxResNet (Chen et al., 2018a), the input and output of each block is summed in a ResNet-style (He et al., 2016). The width of the blocks in the decoder is twice (64) with respect to the encoder part (32) due to the concatenation of previous layers in the same stage of the network.

At the end of the network, the probabilities $\mathbf{z} = [z_1, z_2]$ (corresponding to non-lesion and lesion labels) for each voxel are normalized by the Softmax function

$$q_i = Softmax(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{2} e^{z_j}}, \tag{1}$$

and the segmentation label is $\arg\max_i(q_i), i = 1, 2$.

RatLesNetv2 architecture differs from our previous RatLesNet (Valverde et al., 2019) in two aspects. First, RatLesNetv2 has one additional downsampling and upsampling level, increasing the receptive field to $76 \times 76 \times 76$ voxels. These extra levels allows RatLesNetv2 to consider more information from a larger volume. Second, RatLesNetv2 replaces unpooling (Noh et al., 2015) and DenseNetBlocks (Huang et al., 2017) with trilinear upsampling and ResNetBlocks, respectively, reducing memory usage and execution time. In contrast to VoxResNet (Chen et al., 2018a), RatLesNetv2 architecture resembles an autoencoder, and RatLesNetv2 employs no transposed convolutions, reducing the number of parameters. Additionally, unlike 3D U-Net (Çiçek et al., 2016), RatLesNetv2 uses residual blocks that reutilize previous computed feature maps and facilitate the optimization.

2.4. Loss Function

ConvNets' parameters are optimized by minimizing a loss function that describes the difference between the predictions and the ground truth. RatLesNetv2 was optimized with Adam (Kingma and Ba, 2014) by minimizing cross entropy and Dice loss functions $L_{total} = L_{BCE} + L_{Dice}$. Cross entropy measures the error as the difference between distributions. Since our annotations consist of only two classes (lesion and non-lesion) we used binary cross entropy

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} p_i \cdot \log(q_i) + (1 - p_i) \cdot \log(1 - q_i), \tag{2}$$

where $p_i \in \{0, 1\}$ represents whether voxel i is lesion in the ground truth and $q_i \in [0, 1]$ is the predicted Softmax probability of lesion class. Dice loss (Milletari et al., 2016) is defined as:

$$L_{Dice} = 1 - \frac{2\sum_i^N p_i q_i}{\sum_i^N p_i^2 + \sum_i^N q_i^2}. \tag{3}$$

The rationale behind using Dice loss is to directly maximize the Dice coefficient, one of the metrics to assess image segmentation performance. Although the derivative of Dice loss can be unstable when its denominator is very small, the use of BatchNorm and skip connections helps during the optimization by smoothing the loss landscape (Li et al., 2018; Santurkar et al., 2018).

2.5. Post-processing

Since our model optimizes a per-voxel loss function, small undesirable clusters of voxels may appear disconnected from the main predicted mask. These spurious clusters may be referred as "islands" when they are separated from the largest connected component and "holes" when they are inside the lesion mask. **Figure 1** illustrates these terms.

Small islands and holes can be removed in a final post-processing operation, yielding more realistic segmentations. Determining the maximum size of these holes and islands is, however, challenging in practice: A very small threshold will not eliminate enough small islands and a too large threshold may

remove small lesions. In our experiments, we chose a threshold such that 90% of the holes and islands in the training data were removed. More specifically, we removed holes and islands of 20 voxels or less, inside and outside the lesion masks.

2.6. Evaluation Metrics

We assessed the performance of each ConvNet by measuring the Dice coefficient, Hausdorff distance and compactness. In agreement with the literature (Fenster and Chiu, 2005), we argue that Dice coefficient alone is not an effective measure in rodent lesion segmentation, which is why we complemented it with the two other metrics.

2.6.1. Dice Coefficient

Dice coefficient (Dice, 1945) is one of the most popular metrics in the field of image segmentation. It measures the overlap volume between two binary masks, typically the prediction of the model and the manually-annotated ground truth. Dice coefficient is formally described as:

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{4}$$

where A and B are the segmentation masks.

2.6.2. Compactness

Compact lesion masks are realistic and resemble human-made annotations. Compactness can be defined as the ratio between surface area (*area*) and volume of the mask (*volume*) (Bribiesca, 2008). More specifically, we define compactness as:

$$Compactness = area^{1.5}/volume, \tag{5}$$

which has a constant minimum value of $6\sqrt{\pi}$ for any sphere. Compactness measure penalizes holes, islands and non-smooth borders because these increase the surface area with respect to the volume. Therefore, low compactness values that describe compact segmentations are desirable.

2.6.3. Hausdorff Distance

Hausdorff distance (HD) (Rote, 1991) is defined as:

$$d(A, B) = \max\left\{\max_{a\in\partial A}\min_{b\in\partial B}|b-a|, \max_{b\in\partial B}\min_{a\in\partial A}|a-b|\right\}, \tag{6}$$

where A and B are the segmentation masks, and ∂A and ∂B are their respective boundary voxels. It measures the maximum distance of the ground truth surface to the closest voxel of the prediction, i.e, the largest segmentation error. Measuring Hausdorff distance in brain lesion segmentation studies is crucial since misclassifications far from the lesion boundaries are more severe. The reported Hausdorff distances were in millimeters.

Hausdorff distance and compactness values were calculated exclusively in animals with lesions. Hausdorff distance values on slightly imperfect segmentations of sham-operated animals are excessively large and distort the overall statistics. Additionally, compactness can not be calculated on empty volumes derived from scans without lesions. Voxel anisotropy was accounted for when computing HD and compactness. Finally, we assessed significance of performance difference through a paired permutation test with 10,000 random iterations on the post-processed segmentations with 0.05 as the significance threshold.

2.7. Experimental Setup

2.7.1. Training

RatLesNetv2, 3D U-Net (Çiçek et al., 2016), VoxResNet (Chen et al., 2018a) and RatLesNet (Valverde et al., 2019) were optimized with Adam (Kingma and Ba, 2014) ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$), starting with a learning rate of 10^{-5} for 700 epochs. A small set of learning rates were tested on each architecture to ensure that we used the best performing learning rate in each model. Models were randomly initialized and trained three times separately, and their performance was evaluated from the lesion masks derived with majority voting across these three independent runs. In other words, for each architecture we ensembled three independently trained models. We confirmed that this strategy, typical to remove uncorrelated errors (Dietterich, 2000), improves performance.

2.7.2. Experiments

2.7.2.1. Performance Comparison

We optimized RatLesNetv2, 3D U-Net (Çiçek et al., 2016), VoxResNet (Chen et al., 2018a) and RatLesNet (Valverde et al., 2019) on both the homogeneous and heterogeneous data set divisions (section 2.1) and compared their performance.

2.7.2.2. Ablation Study

We conducted an ablation study (Meyes et al., 2019) in which we changed or removed certain parts of the model to comprehend the effects of the characteristics of RatLesNetv2 architecture. More specifically, we modified the interconnections between layers within each block, changed the number of downsampling/upsampling blocks, and increased and decreased the number of filters.

2.7.2.3. Ground Truth Disparity Effect

We trained two separate RatLesNetv2 models on segmentations annotated by two different operators. This can be seen as an inter-rater variability study of the same ConvNet with disparate knowledge. We run RatLesNetv2 three times for each ground truth on the homogeneous training data, which come exclusively from the study with the two annotations (Study A, **Table 1**). RatLesNetv2 produced six sets of 868 masks $\hat{y}_{g,r}$ where $g \in \{1, 2\}$ refers to the annotator segmenting the training data and $r \in \{1, 2, 3\}$ refers to the run. First, we approximated the intra-rater variability of RatLesNetv2 by calculating the Dice coefficients among the three runs for each ground truth separately, i.e., $\{dice(\hat{y}_{g,1}, \hat{y}_{g,2}), dice(\hat{y}_{g,2}, \hat{y}_{g,3}), dice(\hat{y}_{g,1}, \hat{y}_{g,3})\}$ for $g = 1, 2$. This led to two sets of three Dice coefficients per mask. Second, we calculated the Dice coefficient of the masks across the different ground truths $\{dice(\hat{y}_{1,i}, \hat{y}_{2,j})\}$ for $i, j = 1, 2, 3$ to approximate inter-rater similarity, leading to nine Dice coefficients per mask.

2.7.2.4. Training Set Size

We optimized RatLesNetv2 with training sets of different sizes to understand the relation between training set size and

generalization capability. The training sets had the same ratio of time-points, i.e., we enlarged the training sets by 1 sample per time-point. Since the lowest number of samples across time-points corresponds to 12 (2 h lesions) and we want to keep at least 1 image per time-point in the test set, we produced 11 training sets T_i of size $|T_i| = 8i$ for $i = 1, \ldots, 11$, where 8 is the number of lesion stages.

2.7.3. Implementation

RatLesNetv2 was implemented in Pytorch (Paszke et al., 2019) and it was run on Ubuntu 16.04 with an Intel Xeon W-2125 CPU @ 4.00 GHz processor, 64 GB of memory and an NVidia GeForce GTX 1080 Ti with 11 GB of memory. RatLesNetv2 is publicly available at https://github.com/jmlipman/RatLesNetv2.

3. RESULTS

3.1. Performance of RatLesNetv2

Table 2 lists the quantitative validation results on the test set excluding sham-operated animals that typically yield Dice coefficients of 1.0. As can be seen in **Table 2**, RatLesNetv2 produced similar or better Dice coefficients and Hausdorff distances, and more compact segmentations than the other ConvNets. The average Dice coefficients varied from 0.784 (homogeneous division) to 0.813 (heterogeneous division). Dice coefficients had a large standard deviation regardless of the architecture (from 0.15 to 0.20). However, note that the sample-wise difference between the Dice coefficients of RatLesNetv2 and VoxResNet had a smaller standard deviation of 0.05, i.e., the Dice values between different networks were correlated. **Table 2** shows that RatLesNetv2 achieved significantly better compactness values (all p-values < 0.011) than 3D U-Net, VoxResNet and RatLesNet. Remarkably, 3D U-Net and VoxResNet produced masks with non-smooth borders and several more holes and islands, leading to less compact segmentations (see **Figure 3** and Figures in the **Supplementary Material**). The average compactness values of RatLesNetv2 were higher than the ground truth (20.98 ± 3.28, $p = 0.003$); this was expected as human annotators are likely to produce segmentations with excessively rounded boundaries.

Post-processing had little to no effect on the average Dice coefficients, but it enhanced the final segmentation quality as it removed spurious clusters of voxels. This improvement was reflected in the reduction of compactness values and the considerable decrease of Hausdorff distances. Remarkably, the difference in the Hausdorff distances before and after post-processing was more pronounced in 3D U-Net, VoxResNet and RatLesNet.

Table 3 lists the quantitative results by lesion stage to understand the performance of RatLesNetv2 in detail. Training RatLesNetv2 on the homogeneous data division, whose training set included almost twice as many 24 h lesion scans as the heterogeneous division (9 scans vs. 5 scans), led to a slight increase in the average Dice coefficient and Hausdorff distance in 24 h lesion scans. However, there was no significant difference between either the Dice coefficients ($p = 0.057$) nor Hausdorff distances ($p = 0.08$) of the segmentations derived in the two cases. Dice coefficients, compactness values and Hausdorff distances of the segmentations produced after training on the homogeneous division deteriorated as the time-point was farther from 2 and 24 h.

TABLE 2 | Performance evaluation on the test set before and after post-processing.

Model	Dice (no shams)	Compactness	HD
RatLesNetv2-post	**0.784 ± 0.18**^{*a}	**29.332 ± 7.86**^{*b}	3.522 ± 3.64
RatLesNetv2	0.784 ± 0.18	29.609 ± 8.12	3.687 ± 3.30
3D U-Net-post	0.769 ± 0.20	36.741 ± 11.41	3.665 ± 3.81
3D U-Net	0.768 ± 0.20	37.599 ± 11.77	4.097 ± 3.69
VoxResNet-post	0.757 ± 0.19	37.096 ± 13.00	3.692 ± 3.46
VoxResNet	0.757 ± 0.19	38.161 ± 13.62	4.943 ± 3.38
RatLesNet-post	0.742 ± 0.18	35.045 ± 10.71	3.892 ± 2.54
RatLesNet	0.741 ± 0.18	35.888 ± 10.76	4.679 ± 2.55
RatLesNetv2-post	0.813 ± 0.16	**23.105 ± 4.58**^{*c}	3.334 ± 3.34
RatLesNetv2	0.813 ± 0.16	23.177 ± 4.64	3.512 ± 3.31
3D U-Net-post	0.813 ± 0.15	28.247 ± 5.92	3.099 ± 2.47
3D U-Net	0.812 ± 0.15	28.639 ± 5.99	3.221 ± 2.47
VoxResNet-post	0.806 ± 0.14	32.937 ± 10.05	3.585 ± 3.27
VoxResNet	0.805 ± 0.14	33.634 ± 10.53	4.535 ± 3.46
RatLesNet-post	0.764 ± 0.15	31.348 ± 9.66	3.218 ± 2.79
RatLesNet	0.764 ± 0.15	31.669 ± 9.86	3.384 ± 2.56

Average Dice coefficients were reported in images of animals with lesions. Top: Homogeneous division. Bottom: Heterogeneous division. Bold: Values significantly better than the other architectures ($^{*a}p = 0.007$,$^{*b}p = 0.011$,$^{*c}p = 0.005$).

Training on the heterogeneous training set notably improved the average Dice coefficients and compactness values of every model (**Table 2**) and every time-point (**Table 3**) with respect to homogeneous division, except on 24 h lesions. Furthermore, it decreased the standard deviation of the Dice coefficients and compactness values. RatLesNetv2 recognized animals without lesions notably well even if they were not part of the training set, providing average Dice coefficients of 1.0 on sham-operated animals even without post-processing. Additionally, Dice coefficients on 2 h lesions, 24 h lesions, and overall were higher than inter-rater agreement.

Ensembling three ConvNets of the same architecture optimized on the same training set led to significantly better performance scores in all cases (all p-values < 0.007) as it discarded small segmentation inconsistencies. This strategy increased Dice coefficients by an average of 2% and decreased compactness and Hausdorff distances by an average of 5 and 23% with respect to the first run. The Dice coefficients, compactness values and Hausdorff distances from the individual images used for calculating the reported statistics are also included in the **Supplementary Materials** as CSV files.

3.2. Ablation Studies

The performance scores of RatLesNetv2 after modifying its architecture during the ablation studies are reported in **Table 4**.

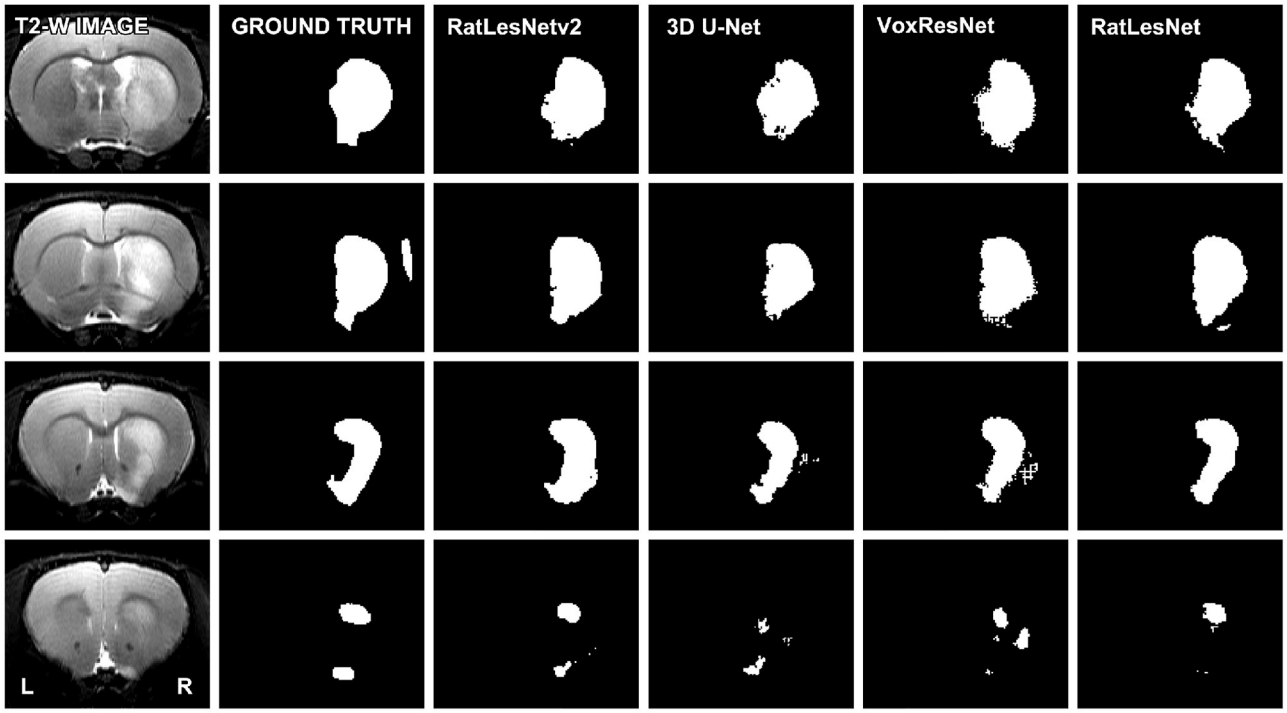


FIGURE 3 | Comparison of the segmentation masks of four consecutive slices. The depicted T2-weighted image corresponds to a typical scan, i.e., the volume whose segmentation achieved the median Dice coefficient in the test set (heterogeneous division). Segmentations were not post-processed.

3.2.1. DenseNetBlock

Similarly to RatLesNet (Valverde et al., 2019), DenseNet-style (Huang et al., 2017) blocks were implemented in RatLesNetv2 while keeping the same number of parameters of the baseline RatLesNetv2 model. Dice coefficients and compactness values were significantly deteriorated with respect to RatLesNetv2 baseline (all p-values < 0.037), and Hausdorff distances increased slightly in homogeneous data division, whereas they decreased in heterogeneous division. Additionally, DenseNetBlocks demanded notably more memory due to the concatenation operation.

3.2.2. Halving the Receptive Field (RF)

The third downsampling stage of RatLesNetv2 was eliminated in order to reduce the receptive field from 72 voxels down to 36. An additional test (marked in **Table 4** with an *) matched the number of parameters to the baseline. The reduction of the receptive field yielded in significant improvements of the Dice coefficient and a significant deterioration of the compactness and Hausdorff distance in the heterogeneous division (all p-values < 0.028). On the other hand, in the homogeneous division Dice coefficients and compactness values were worse than RatLesNetv2 baseline.

3.2.3. Network Width

We increased and decreased the number of filters of RatLesNetv2 by 4 (**Table 4**, Width-28 and Width-36). This modification decreased the Dice coefficients with respect to RatLesNetv2 and led to no significant difference in the Hausdorff distances. Compactness values showed contradictory results; they deteriorated in homogeneous division whereas they remained similar or slightly worse in heterogeneous division.

3.3. On the Influence of Disparate Ground Truths

As expected, optimizing separate RatLesNetv2 models with segmentations from different annotators produced more different segmentation masks than when optimizing with segmentations from the same annotator. In other words, the three sets of predictions $\hat{y}_{1,1}, \hat{y}_{1,2}, \hat{y}_{1,3}$ were similar among themselves in the same manner as $\hat{y}_{2,1}, \hat{y}_{2,2}, \hat{y}_{2,3}$ (**Figure 4B**, Annotation 1 and 2), and their differences arise from the stochasticity of ConvNets optimization. In contrast, the shape of the distribution of the Dice coefficients that compare masks derived from RatLesNetv2 models optimized with different annotations (**Figure 4B**, Mixed) was notably different. Also, Annotation 1 and Mixed Dice coefficients as well as Annotation 2 and Mixed Dice coefficients were significantly different (p-values < 0.002).

In a visual inspection, we observed that Annotation 2 was more approximate, with simpler contours, than Annotation 1. **Figure 4A** (top row) shows the manual segmentations of the scan with the most disparate annotations and **Figure 4A** (bottom row) shows the predictions on a scan with the highest Dice coefficient on our baseline study when RatLesNetv2 was trained on the different annotations.

TABLE 3 | Performance evaluation on the test set after post-processing segregated by lesion stage.

Time-point (scans)	Dice	Compactness	HD
24 h (394)	0.831 ± 0.15	26.539 ± 4.86	3.691 ± 3.53
D3 (135)	0.782 ± 0.12	29.705 ± 8.01	3.067 ± 2.31
D7 (53)	0.790 ± 0.11	40.742 ± 10.65	2.580 ± 2.63
D14 (30)	0.735 ± 0.21	36.018 ± 11.07	3.329 ± 5.06
D21 (40)	0.800 ± 0.11	33.797 ± 6.50	2.546 ± 0.92
D28 (89)	0.593 ± 0.28	29.598 ± 7.46	4.238 ± 5.22
D35 (20)	0.751 ± 0.23	31.203 ± 4.72	4.831 ± 5.76
Shams (107)	1.000 ± 0.00	—	—
2 h (6)	0.719 ± 0.11	23.111 ± 2.27	1.920 ± 0.16
24 h (400)	0.826 ± 0.15	23.218 ± 4.67	3.919 ± 3.79
D3 (129)	0.809 ± 0.10	23.376 ± 5.15	2.796 ± 2.24
D7 (47)	0.860 ± 0.09	23.555 ± 3.99	2.439 ± 2.83
D14 (24)	0.827 ± 0.19	21.705 ± 3.36	3.015 ± 5.69
D21 (34)	0.877 ± 0.10	23.874 ± 2.38	2.002 ± 0.70
D28 (83)	0.692 ± 0.25	22.147 ± 4.65	2.875 ± 1.93
D35 (14)	0.886 ± 0.07	22.037 ± 2.55	1.700 ± 0.67
Shams (131)	1.000 ± 0.00	—	—

Top: Homogeneous division. Bottom: Heterogeneous division.

TABLE 4 | Ablation study.

Study	Dice (no shams)	Compactness	HD
Baseline	0.784 ± 0.18	29.332 ± 7.86	3.522 ± 3.64
DenseNetBlock*	**0.771 ± 0.20**	**30.094 ± 8.86**	3.692 ± 3.96
Halving RF	**0.754 ± 0.20**	30.766 ± 10.57	3.340 ± 3.91
Halving RF*	**0.765 ± 0.19**	**31.867 ± 10.41**	3.464 ± 3.53
Width-28	0.781 ± 0.18	30.095 ± 8.42	3.423 ± 2.84
Width-36	**0.765 ± 0.19**	**31.620 ± 10.09**	3.557 ± 3.73
Baseline	0.813 ± 0.16	23.105 ± 4.58	3.334 ± 3.34
DenseNetBlock*	**0.801 ± 0.16**	**23.313 ± 5.13**	3.093 ± 2.70
Halving RF	*0.819 ± 0.15*	**25.226 ± 5.34**	**3.679 ± 3.12**
Halving RF*	*0.820 ± 0.15*	**25.394 ± 5.51**	**3.719 ± 3.40**
Width-28	**0.803 ± 0.17**	22.861 ± 4.63	*2.892 ± 2.94*
Width-36	**0.801 ± 0.16**	24.036 ± 5.04	2.900 ± 2.84

*Top: Homogeneous task. Bottom: Heterogeneous task. Bold: baseline significantly better. Italic: baseline significantly worse (p-values < 0.05). *Equal number of parameters as Baseline.*

3.4. The Impact of the Training Set Size on the Performance

Figure 5 illustrates the evolution of the Dice coefficients, compactness values and Hausdorff distances as the training set increases in size. Dice coefficients (**Figure 5**, left) were remarkably different across time-points and almost every time-point reached a performance plateau with large data sets. Time-points 24 h and D3—which composed the majority of the test set scans by 56.7 and 17.8% of the total, respectively—reached their plateaus later. This effect can be a consequence of the variability within samples. On the contrary, the time-points with the lowest number of samples (2 h and D35 lesions with 1 and 9 image, respectively) exhibited fluctuations.

Compactness values (**Figure 5**, center) and Hausdorff distances (**Figure 5**, right) oscillated considerably regardless of the time-point. Hausdorff distances were higher in the time-points with the largest number of samples (24 h and D3), likely due to the existence of outliers. Compactness values, including the average (dashed line), increased analogously to the training set size, i.e., enlarging the training set yielded less compact segmentations. Yet, these compactness values were markedly lower than the compactness values derived from segmentations produced by 3D U-Net, VoxResNet, and RatLesNet (section 3.1).

4. DISCUSSION

We showed that RatLesNetv2 yielded similar or better Dice coefficients and Hausdorff distances, and notably more compact segmentations than other convolutional neural networks (Çiçek et al., 2016; Chen et al., 2018a; Valverde et al., 2019). These measurements indicate that the segmentations derived from RatLesNetv2 were more similar to the ground truth, had less large segmentation errors and were more realistic. Additionally, the smaller differences between Hausdorff distances before and after post-processing derived from RatLesNetv2 also indicate that RatLesNetv2 produced fewer segmentation errors far from the lesion surface.

RatLesNetv2 produced more compact segmentations than the other ConvNets without directly minimizing compactness (see **Table 2**), indicating that RatLesNetv2 architecture favors segmentations with smooth borders without holes. Although optimizing compactness (and Hausdorff distance) directly might further improve the results, incorporating these terms to the loss function leads to additional hyper-parameters that require costly tuning. Dice coefficients had large standard deviations and were lower than in existing human brain tumor segmentation studies (Jiang et al., 2019; Myronenko and Hatamizadeh, 2019). These results may arise due to the subjectivity of the segmentation task caused by low image contrast in certain lesions and its consequent high inter- and intra-rater disagreement. However, this is not unexpected as relatively low Dice coefficients and large standard deviations are typical in rodent (Mulder et al., 2017a; Valverde et al., 2019) and human brain lesion segmentation studies (Chen et al., 2017; Valverde et al., 2017; Subbanna et al., 2019), even when studying inter-rater disagreement of manual annotations relying on a semi-automatic segmentation pipeline (Mulder et al., 2017a). We also argued that Dice coefficient alone is not sufficient to measure the segmentation performance. To illustrate the importance of providing additional measurements, consider a brain with a very large and a very small lesion. If the segmentation accurately predicts the large lesion and ignores the small one, Dice coefficients will have a high value not reflecting the segmentation error, but Hausdorff distance is high capturing the segmentation error. Likewise, a lesion segmentation mask with non-smooth surface and several small holes and islands

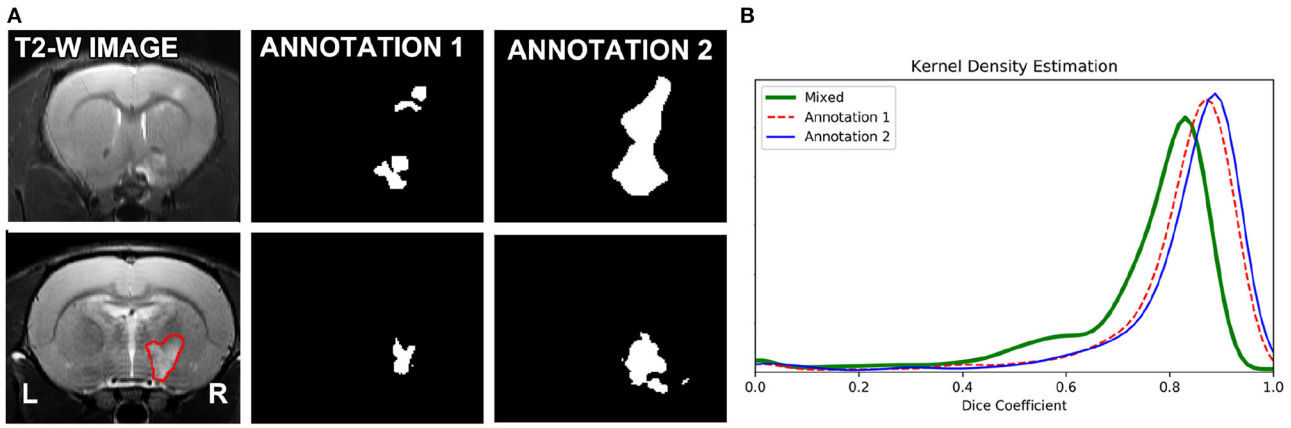


FIGURE 4 | (A) (Top row): Scan with the most disparate annotations between operators 1 and 2. **(A)** (Bottom row): A randomly selected scan of the test set (left), segmentations of the scan with RatLesNetv2 trained on Annotator 1 ground truth (middle), and Annotator 2 ground truth (right). **(B)** Kernel density estimation of three sets of Dice coefficients. Red (dashed line) and blue (solid line) estimations were calculated between the predictions of the model optimized for the same ground truth. Green (thick solid line) estimation was computed between the predictions whose model was optimized for different ground truths. The predictions generated when the same model is optimized for different ground truths are notably different.

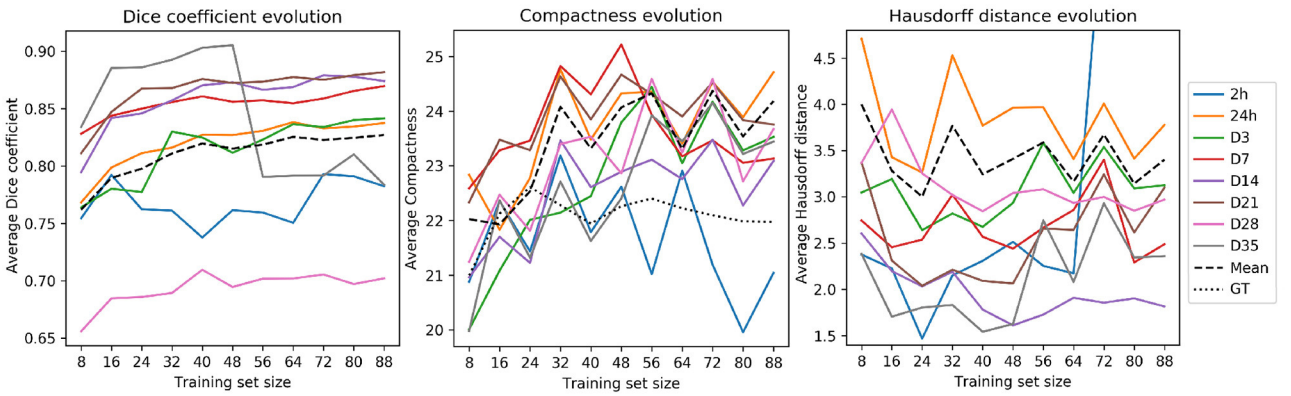


FIGURE 5 | RatLesNetv2 performance when optimizing for training sets of multiple sizes. Metrics (from left to right: Dice coefficient, compactness and Hausdorff distance) were processed from the masks derived with the majority voting across three runs on a fixed test set (828 images). Averages (dashed lines) were segregated by time-point. Compactness graph includes the average compactness of the ground truth (dotted line).

(i.e., a high compactness value) may have a high Dice coefficient despite being unrealistic.

The difference in the performance between homogeneous and heterogeneous data set divisions indicates that although few 24 h lesion volumes were needed to generalize well, adding more 24 h lesion volumes to the training data (homogeneous division) made RatLesNetv2 specialize on that time-point (**Table 3**). On the other hand, increasing data diversity (heterogeneous division) improved performance, demonstrating that RatLesNetv2 is capable of learning from a heterogeneous data set. Thus, training on this heterogeneous division increased RatLesNetv2 capability to extrapolate to different-looking ischemic brain lesions. However, without optimizing on additional data, RatLesNetv2 performance on images with other types of lesions, such as tumor lesions, is limited by the lesions' appearance.

The ablation experiments showed that modifications of RatLesNetv2 architecture yielded similar or worse performance, justifying RatLesNetv2's architectural choices. Despite both residual connections (He et al., 2016) and DenseNetBlocks (Huang et al., 2017) facilitate gradient propagation (Drozdzal et al., 2016), residual connections were preferred over DenseNetBlocks due to their notably higher performance and lower memory requirements. Additionally, a large receptive field empirically demonstrated to increase compactness and reduce large segmentation errors possibly because RatLesNetv2 considers a larger context. The choice of a large receptive field is in agreement with other state-of-the-art ConvNets that achieve large receptive fields by stacking several convolutional layers and/or utilizing dilated convolutions (Chen et al., 2018b).

Our ground-truth disparity experiment confirmed that predictions generated when the same model is optimized for different ground truths are different. Consequently, the quality of the manually-annotated ground truth has a direct impact on the quality of the lesion masks generated automatically. As there

is no unique definition of "lesion," it may be advantageous for an algorithm to perform differently depending on the labels of the training set. On the other hand, it may also be desirable to design a robust algorithm that performs consistently regardless of some changes in the annotations.

The experiment of training RatLesNetv2 on several training sets of different sizes showed that even with few available training data RatLesNetv2 can generalize well and, despite increasing its performance when optimizing on larger training sets, such improvement is small and compactness values and Hausdorff distances fluctuate considerably.

5. CONCLUSION

We presented and made publicly available RatLesNetv2, a 3D ConvNet to segment rodent brain lesions. RatLesNetv2 has been evaluated on an exceptionally large and diverse data set of 916 rat brain MR images, validating RatLesNetv2 reliability on a wide variety of lesion stages with lesions of different appearance. Additionally, RatLesNetv2 produced segmentations that exceeded overall inter-rater agreement Dice coefficients (inter-rater: 0.73 ± 0.12, RatLesNetv2: 0.81 ± 0.16). This enhancement indicates that RatLesNetv2 produces segmentations that are remarkably more consistent with the ground truth than the similarity between different human-made annotations. This consistency is of special importance for research reproducibility, crucial in preclinical studies.

Based on our experiments and, more specifically, the accuracy greater than inter-rater agreement and than of other ConvNets, RatLesNetv2 can be used to automate lesion segmentation in preclinical MRI studies on rats.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: the software and the trained models are publicly available from https://github.com/jmlipman/RatLesNetv2. All the quantitative segmentation measures are **Supplementary Material** of this manuscript. The MR image dataset cannot be shared due to ownership and intellectual property restrictions.

ETHICS STATEMENT

The animal study was reviewed and approved by National Animal Experiment Board, Finland.

AUTHOR CONTRIBUTIONS

JV and JT designed the study and analyzed the data. JV conceptualized the algorithm design, developed RatLesNetv2, and wrote the draft. JV, RD, and JT developed the methodology. ASh collected the data. ASh, OG, and ASi interpreted the data. All authors contributed to manuscript revision, proofreading, and approved the submitted version.

FUNDING

The work of JV was funded from the European Union's Horizon 2020 Framework Programme [Marie Skłodowska Curie grant agreement #740264 (GENOMMED)] and RD's work was funded from Marie Skłodowska Curie grant agreement #691110 (MICROBRADAM). We also acknowledge the Academy of Finland grants (#275453 to ASi and #316258 to JT).

ACKNOWLEDGMENTS

Part of the computational analysis was run on the servers provided by Bioinformatics Center, University of Eastern Finland, Finland. This manuscript has been released as a pre-print at arXiv (Valverde et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.610239/full#supplementary-material

REFERENCES

Arnaud, A., Forbes, F., Coquery, N., Collomb, N., Lemasson, B., and Barbier, E. L. (2018). Fully automatic lesion localization and characterization: application to brain tumors using multiparametric quantitative mri data. *IEEE Trans. Med. Imaging* 37, 1678–1689. doi: 10.1109/TMI.2018.2794918

Bribiesca, E. (2008). An easy measure of compactness for 2D and 3D shapes. *Pattern Recogn.* 41, 543–554. doi: 10.1016/j.patcog.2007.06.029

Chen, H., Dou, Q., Yu, L., Qin, J., and Heng, P.-A. (2018a). Voxresnet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage* 170, 446–455. doi: 10.1016/j.neuroimage.2017.04.041

Chen, L., Bentley, P., and Rueckert, D. (2017). Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *Neuroimage Clin.* 15, 633–643. doi: 10.1016/j.nicl.2017.06.016

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 801–818. doi: 10.1007/978-3-030-01234-2_49

Choi, C.-H., Yi, K. S., Lee, S.-R., Lee, Y., Jeon, C.-Y., Hwang, J., et al. (2018). A novel voxel-wise lesion segmentation technique on 3.0-T diffusion MRI of hyperacute focal cerebral ischemia at 1 h after permanent MCAO in rats. *J. Cereb. Blood Flow Metab.* 38, 1371–1383. doi: 10.1177/0271678X17714179

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

De Feo, R., and Giove, F. (2019). Towards an efficient segmentation of small rodents brain: a short critical review. *J. Neurosci. Methods* 323, 82–89. doi: 10.1016/j.jneumeth.2019.05.003

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

Dietterich, T. G. (2000). "Ensemble methods in machine learning," in *International Workshop on Multiple Classifier Systems* (Cagliari: Springer), 1–15. doi: 10.1007/3-540-45014-9_1

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, eds G. Carneiro, D. Mateus, P. Loïc, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, and J. Cornebise (Athens: Springer), 179–187. doi: 10.1007/978-3-319-46976-8_19

Duong, M. T., Rudie, J. D., Wang, J., Xie, L., Mohan, S., Gee, J. C., et al. (2019). Convolutional neural network for automated flair lesion segmentation on clinical brain mr imaging. *Am. J. Neuroradiol*. 40, 1282–1290. doi: 10.3174/ajnr.A6138

Dutta, S., and Sengupta, P. (2016). Men and mice: relating their ages. *Life Sci*. 152, 244–248. doi: 10.1016/j.lfs.2015.10.025

Fenster, A., and Chiu, B. (2005). "Evaluation of segmentation algorithms for medical imaging," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* (Shanghai), 7186–7189. doi: 10.1109/IEMBS.2005.1616166

Gabr, R. E., Coronado, I., Robinson, M., Sujit, S. J., Datta, S., Sun, X., et al. (2019). Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: a large-scale study. *Mult. Scler. J*. 26, 1217–1226. doi: 10.1177/1352458519856843

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 4700–4708. doi: 10.1109/CVPR.2017.243

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv* 1502.03167.

Jiang, Z., Ding, C., Liu, M., and Tao, D. (2019). "Two-stage cascaded U-net: 1st place solution to brats challenge 2019 segmentation task," in *International MICCAI Brainlesion Workshop* (Shenzhen: Springer), 231–241. doi: 10.1007/978-3-030-46640-4_22

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *CoRR* abs/1412.6980.

Koizumi, J., Yoshida, Y., Nakazawa, T., and Ooneda, G. (1986). Experimental studies of ischemic brain edema. 1. A new experimental model of cerebral embolism in rats in which recirculation can be introduced in the ischemic area. *Jpn. J. Stroke* 8, 1–8. doi: 10.3995/jstroke.8.1

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC), 6389–6399.

Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv* abs/1901.08644.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571. doi: 10.1109/3DV.2016.79

Moraga, A., Gómez-Vallejo, V., Cuartero, M. I., Szczupak, B., San Sebastián, E., Markuerkiaga, I., et al. (2016). Imaging the role of toll-like receptor 4 on cell proliferation and inflammation after cerebral ischemia by positron emission tomography. *J. Cereb. Blood Flow Metab*. 36, 702–708. doi: 10.1177/0271678X15627657

Mulder, I. A., Khmelinskii, A., Dzyubachyk, O., de Jong, S., Rieff, N., Wermer, M. J., et al. (2017a). Automated ischemic lesion segmentation in mri mouse brain data after transient middle cerebral artery occlusion. *Front. Neuroinform*. 11:3. doi: 10.3389/fninf.2017.00003

Mulder, I. A., Khmelinskii, A., Dzyubachyk, O., De Jong, S., Wermer, M. J., Hoehn, M., et al. (2017b). MRI mouse brain data of ischemic lesion after transient middle cerebral artery occlusion. *Front. Neuroinform*. 11:51. doi: 10.3389/fninf.2017.00051

Myronenko, A., and Hatamizadeh, A. (2019). "Robust semantic segmentation of brain tumor regions from 3D MRIs," in *International MICCAI Brainlesion Workshop* (Shenzhen: Springer), 82–89. doi: 10.1007/978-3-030-46643-5_8

Noh, H., Hong, S., and Han, B. (2015). "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago), 1520–1528. doi: 10.1109/ICCV.2015.178

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC), 8024–8035.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Rote, G. (1991). Computing the minimum hausdorff distance between two point sets on a line under translation. *Inform. Process. Lett*. 38, 123–127. doi: 10.1016/0020-0190(91)90233-8

Roy, S., Knutsen, A., Korotcov, A., Bosomtwi, A., Dardzinski, B., Butman, J. A., et al. (2018). "A deep learning framework for brain extraction in humans and animals with traumatic brain injury," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (Washington, DC: IEEE), 687–691. doi: 10.1109/ISBI.2018.8363667

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018). "How does batch normalization help optimization?" in *Advances in Neural Information Processing Systems*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC), 2483–2493.

Subbanna, N. K., Rajashekar, D., Cheng, B., Thomalla, G., Fiehler, J., Arbel, T., et al. (2019). Stroke lesion segmentation in flair MRI datasets using customized markov random fields. *Front. Neurol*. 10:541. doi: 10.3389/fneur.2019.00541

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2818–2826. doi: 10.1109/CVPR.2016.308

Valverde, J. M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A., and Tohka, J. (2019). "Automatic rodent brain mri lesion segmentation with fully convolutional networks," in *International Workshop on Machine Learning in Medical Imaging* (Shenzhen: Springer), 195–202. doi: 10.1007/978-3-030-32692-0_23

Valverde, J. M., Shatillo, A., De Feo, R., Gröhn, O., Sierra, A., and Tohka, J. (2020). Ratlesnetv2: a fully convolutional network for rodent brain lesion segmentation. *arXiv* 2001.09138.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., et al. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* 155:159–168. doi: 10.1016/j.neuroimage.2017.04.034

Wang, Y., Cheung, P.-T., Shen, G. X., Bhatia, I., Wu, E. X., Qiu, D., et al. (2007). Comparing diffusion-weighted and T2-weighted mr imaging for the quantification of infarct size in a neonatal rat hypoxic-ischemic model at 24 h post-injury. *Int. J. Deve. Neurosci*. 25, 1–5. doi: 10.1016/j.ijdevneu.2006.12.003

Yang, H., Huang, W., Qi, K., Li, C., Liu, X., Wang, M., et al. (2019). "CLCI-net: cross-level fusion and context inference networks for lesion segmentation of chronic stroke," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen: Springer), 266–274. doi: 10.1007/978-3-030-32248-9_30

Conflict of Interest: As disclosed in the affiliation section, ASh is a full-time payroll employee of the Charles River Discovery Services, Finland—a commercial pre-clinical contract research organization (CRO), which participated in the project and provided raw data as a part of company's R&D initiative.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neurology

ORIGINAL RESEARCH
published: 22 February 2021
doi: 10.3389/fneur.2021.613967

Diagnostic Performance of MRI Volumetry in Epilepsy Patients With Hippocampal Sclerosis Supported Through a Random Forest Automatic Classification Algorithm

Juan Pablo Princich[1,2], Patricio Andres Donnelly-Kehoe[3], Alvaro Deleglise[4], Mariana Nahir Vallejo-Azar[1], Guido Orlando Pascariello[3], Pablo Seoane[1,5], Jose Gabriel Veron Do Santos[1], Santiago Collavini[1,6,7], Alejandro Hugo Nasimbera[1,5] and Silvia Kochen[1]*

[1] ENyS (Estudios en Neurociencias y Sistemas Complejos), Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional Arturo Jauretche y Hospital El Cruce, Florencio Varela, Argentina, [2] Hospital de Pediatría J.P Garrahan, Departamento de Neuroimágenes, Buenos Aires, Argentina, [3] Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CIFASIS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Grupo de Procesamiento de Señales Multimedia - División Neuroimágenes, Universidad Nacional de Rosario, Rosario, Argentina, [4] Instituto de Fisiología y Biofísica B. Houssay (IFIBIO), Consejo Nacional de Investigaciones Científicas y Técnicas, Departamento de Fisiología y Biofísica, Universidad de Buenos Aires, Buenos Aires, Argentina, [5] Hospital J.M Ramos Mejía, Centro de Epilepsia, Buenos Aires, Argentina, [6] Instituto de investigación en Electrónica, Control y Procesamiento de Señales (LEICI), Universidad Nacional de La Plata-Consejo Nacional de Investigaciones Científicas y Técnicas, La Plata, Argentina, [7] Instituto de Ingeniería y Agronomía, Universidad Nacional Arturo Jauretche, Florencio Varela, Argentina

OPEN ACCESS

Edited by:
Hans-Peter Müller, University of Ulm, Germany

Reviewed by:
Jordi A. Matias-Guiu, Hospital Clínico San Carlos, Spain
Alessia Sarica, University of Magna Graecia, Italy

***Correspondence:**
Juan Pablo Princich
jpprincich@gmail.com

Specialty section:
This article was submitted to Applied Neuroimaging, a section of the journal Frontiers in Neurology

Received: *05 October 2020*
Accepted: *18 January 2021*
Published: *22 February 2021*

Citation:
Princich JP, Donnelly-Kehoe PA, Deleglise A, Vallejo-Azar MN, Pascariello GO, Seoane P, Veron Do Santos JG, Collavini S, Nasimbera AH and Kochen S (2021) Diagnostic Performance of MRI Volumetry in Epilepsy Patients With Hippocampal Sclerosis Supported Through a Random Forest Automatic Classification Algorithm. Front. Neurol. 12:613967. doi: 10.3389/fneur.2021.613967

Introduction: Several methods offer free volumetry services for MR data that adequately quantify volume differences in the hippocampus and its subregions. These methods are frequently used to assist in clinical diagnosis of suspected hippocampal sclerosis in temporal lobe epilepsy. A strong association between severity of histopathological anomalies and hippocampal volumes was reported using MR volumetry with a higher diagnostic yield than visual examination alone. Interpretation of volumetry results is challenging due to inherent methodological differences and to the reported variability of hippocampal volume. Furthermore, normal morphometric differences are recognized in diverse populations that may need consideration. To address this concern, we highlighted procedural discrepancies including atlas definition and computation of total intracranial volume that may impact volumetry results. We aimed to quantify diagnostic performance and to propose reference values for hippocampal volume from two well-established techniques: FreeSurfer v.06 and volBrain-HIPS.

Methods: Volumetry measures were calculated using clinical T1 MRI from a local population of 61 healthy controls and 57 epilepsy patients with confirmed unilateral hippocampal sclerosis. We further validated the results by a state-of-the-art machine learning classification algorithm (Random Forest) computing accuracy and feature relevance to distinguish between patients and controls. This validation process was performed using the FreeSurfer dataset alone, considering morphometric values not only from the hippocampus but also from additional non-hippocampal brain regions that could be potentially relevant for group classification. Mean

reference values and 95% confidence intervals were calculated for left and right hippocampi along with hippocampal asymmetry degree to test diagnostic accuracy.

Results: Both methods showed excellent classification performance (AUC:> 0.914) with noticeable differences in absolute (cm^3) and normalized volumes. Hippocampal asymmetry was the most accurate discriminator from all estimates (AUC:1∼0.97). Similar results were achieved in the validation test with an automatic classifier (AUC:>0.960), disclosing hippocampal structures as the most relevant features for group differentiation among other brain regions.

Conclusion: We calculated reference volumetry values from two commonly used methods to accurately identify patients with temporal epilepsy and hippocampal sclerosis. Validation with an automatic classifier confirmed the principal role of the hippocampus and its subregions for diagnosis.

Keywords: epilepsy, volumetry, hippocampal sclerosis, random forest classifier, MRI

INTRODUCTION

Quantification of brain anatomical structures from magnetic resonance images (MR) is being increasingly used to recognize pathologic conditions such as temporal lobe epilepsy. Volumetric estimates of hippocampal size are postulated to be more sensitive than visual assessment alone, and also to improve clinical diagnosis in dementia and epilepsy (1–4).

Temporal lobe epilepsy with hippocampal sclerosis (HS) is one of the most frequent focal epilepsies in adults often refractory to pharmacological treatment; surgical resection is an effective therapeutic option for these patients achieving a seizure-free rate close to 80%.

Patients with temporal epilepsy and HS usually share clinical key features associated with the majority of seizure discharges including characteristic aura, arrest, alteration of consciousness (and amnesia), and automatisms. Relatively typical scalp EEG findings can be seen in the interictal state, at the seizure onset, during the course of the seizure, and postictally.

Hippocampal sclerosis is suspected in epilepsy patients when compatible ictal semiology and scalp EEG findings are found, but definitive diagnosis is established based on characteristic brain MR anomalies. Neuroimaging abnormalities are typically recognized in the hippocampus proper, including atrophy, loss of internal structure, and decreased T1- and increased T2-FLAIR signal intensity in clinical practice (5). Inspection of hippocampal coronal sections allows for a side-by-side comparison of asymmetry in volume, shape, and signal important for clinical diagnosis. Atrophy seems to be the most specific and signal changes the most sensitive biomarker in HS (6). Magnets with high field strengths above 3 T are able to depict subtle blurring of the internal architecture of the hippocampus on T2-weighted images (5). Originally, manual segmentation of the hippocampus based on anatomical knowledge and specific MRI landmarks was used to estimate structural volumes. Previous studies using these methods adequately identified lateralization of seizure origin in the temporal lobe of patients with HS. Earlier reports also documented a strong association between severity of histopathological anomalies and hippocampal volumes with an increased diagnostic yield of MR studies (7, 8)

The recent development of automatic volumetry methods such as FreeSurfer (FS) suite (9) and VolBrain (vB) HIPS (10), among others, makes it possible to account for hippocampal volume differences that may escape visual detection. Several studies validated the utility of hippocampal volumetry for HS detection in temporal epilepsy, mostly based on postoperative correlation or using *ex vivo* neuroimaging analysis (7–10). The potential of volumetry measures for postsurgical outcome prediction is still modest, with some improvement in reports considering subfields patterns of atrophy (11).

Since numerous publications demonstrate considerable differences in normative brain structural volumes across populations with different genetic backgrounds (12–14), volumetric estimates of the hippocampus in different populations are of particular concern. Previous reports consistently show hippocampal volume differences even when using analogous procedures (15–32). In this regard, normal anatomic variations and differences associated with the implemented methodology need to be considered for the interpretations of clinical conditions. An additional concern is that normative structural data from Latin America populations remains underrepresented.

The main objective of this work is to estimate reference values of sensitivity, specificity, and confidence intervals for classification of a local population of epilepsy patients with unilateral hippocampal sclerosis using two different volumetry approaches.

We analyzed T1 brain MRI volumetry of the hippocampus and hippocampal subfields in a cohort of 61 healthy subjects and in 57 epilepsy patients with confirmed unilateral mesial temporal sclerosis. Anatomical volumes were computed using two well-established automatic methods FS and vB. Recorded values for the hippocampus and subregions are expressed as absolute values (in cm^3) and further normalized to brain size, quantified as a percent of total intracranial volume (TIV).

Furthermore, we provide hippocampal and subfield volume distribution for a community-based sample of healthy controls

(HC) and evaluate subregion asymmetry differences in HC and between patients. We also compared the degree of asymmetry in left and right HS to investigate its relevance for diagnosis and the presence of distinctive patterns of atrophy at the subregion level.

Finally, a validation process was implemented to explore the contribution of non-hippocampal structures for group classification. This was performed using machine learning techniques, considering only FreeSurfer's morphometric information of whole-brain regions, including anatomical volumes and cortical thickness. Specifically, we used a feature selection technique to obtain the optimal number of features to discriminate between patients and HC, and then we performed three binary classifications for each group using a Monte Carlo cross-validation (MCCV) scheme (33) with a random forest classifier.

MATERIALS AND METHODS

Participants

Patients were retrospectively enrolled based on medical records from the epilepsy unit between 2014 and 2019 at Nestor Kirchner—El Cruce Hospital at Florencio Varela, Buenos Aires, with a final diagnosis of temporal lobe epilepsy associated with unilateral right ($n = 22$, 15 females) and left ($n = 35$, 17 females) hippocampal sclerosis. Diagnosis was established using standardized practices as described in Oddo et al. (34) through clinical examination, assessment of disease history, semiology of seizures along with neuropsychological tests including prolonged video EEG, and compatible findings on 3-T MRI as suggested by ILAE (5). Thirty-one patients (54%) underwent surgical treatment with histopathology confirmation of HS after standard amygdalohippocampectomy with partial temporal lobectomy. The remaining patients are not yet operated but scheduled for surgery. Age- and sex-paired HC ($n = 61$, 44 females) were recruited mostly from local universities including students and academic personnel.

All participants gave written consent to participate and to make use of medical information for this study. The work described in this paper was carried out in accordance with the code of ethics of the world medical association (Declaration of Helsinki). Research ethics approval was obtained from the Hospital Research Ethics Board at El Cruce Hospital.

Imaging Characteristics and Analysis Methods

Only volumetric T1-weighted images were used in this study. These images were obtained as part of the clinical protocol for epilepsy workout in our institution and were acquired using the same MR unit (Philips Achieva 3T, 8-channel head coil), as recommended on recent specialized guidelines (5). Structural images consist of a 3D T1WI (FFE) sequence, with 180 slices of 1-mm isotropic resolution, TE= 3.3 msec, TR= 2300 msec, TI= 900 msec, flip angle= 9°, and field of view (FOV)= 240 × 240 × 180. Images were exported from the scanner and transformed to Nifti format for further analysis. For the statistical analysis, the same T1 volumetric images were processed using two established and freely available methods used to calculate brain region segmentation and quantification, namely, FreeSurfer Suite v6.0 (FS) working in an offline workstation and VolBrain-Hips 2016 (vB) that provides online services running on remote servers through a website interface.

Both methods offer validated hippocampal and hippocampal subfield segmentation through different approaches, distinct reference atlases, dissimilar processing times, and specific subfield region delineations. Output files and results from both methods were independently reviewed by two experienced neuroradiologists (JPP and GDS) looking for labeling inconsistencies and to assure quality control (no manual correction was performed). (See segmentation details for each method in **Figure 1**). Full documentation is available for processing details on each software platform, but here we describe a resumed version of each method.

FREESURFER V6.0

All T1 brain volumes were processed to obtain a complete morphometric description. Cortical reconstruction and volumetric segmentation were performed in each participant's native space on FreeSurfer's[1] (v 6.0) image analysis suite.

Briefly, image processing included removal of non-brain tissue using a hybrid watershed/surface deformation procedure, an automatic Talairach transformation, segmentation of the subcortical WM and deep GM volumetric structures (including hippocampus, amygdala, caudate, putamen, and ventricles), intensity normalization, tessellation of the GM–WM boundary, an automatic topology correction, and surface deformation following intensity gradients to optimally place the GM/WM and GM/CSF borders at the location where the greatest shift in intensity defines the transition to the other tissue class (9).

Once the cortical models were complete, a number of deformable procedures were performed for further data processing and analysis, including surface inflation and registration to a spherical atlas—based on individual cortical folding patterns to match cortical geometry across subjects, parcellation of the cerebral cortex into units relative to gyral and sulcal structure, and creation of a variety of surface-based data—including maps of curvature and sulcal depth. These methods use both intensity and continuity information of the entire 3D MR volume from segmentation and deformation procedures to produce representations of cortical thickness, which is calculated as the closest distance from the GM/WM boundary to the GM/CSF boundary at each vertex on the tessellated surface (9). The maps were created using spatial intensity gradients across tissue classes; therefore, they were not simply reliant on absolute signal intensity. Since the ensuing maps were not restricted to the voxel resolution of the original data, they can detect submillimeter differences between groups. Procedures for the measurement of cortical thickness have been validated against histological analysis and manual measurements. FreeSurfer morphometric procedures including principal hippocampal subfields have been demonstrated to

[1]https://surfer.nmr.mgh.harvard.edu/
[2]https://www.slicer.org

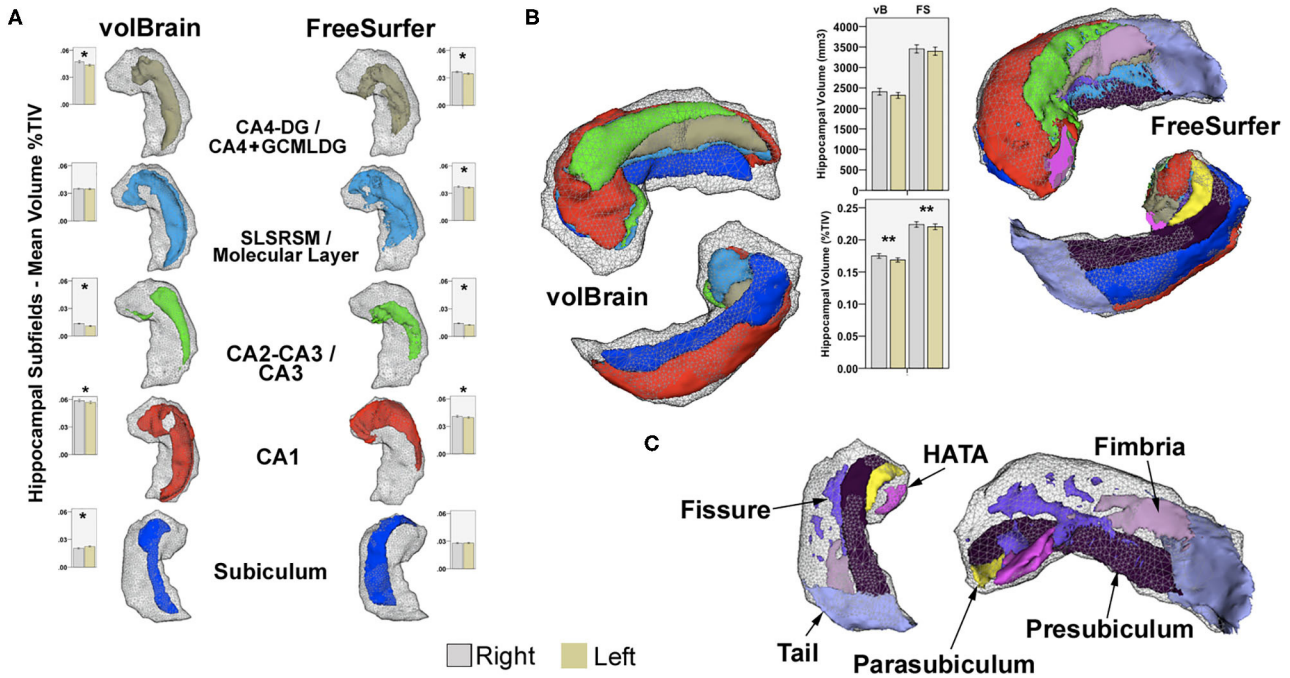


FIGURE 1 | (A) Examples of common subfield's atlas definition using vB and FS. Boxplots represent mean volumes as percent of TIV and whiskers the 95% confidence interval for the HC group. **(B)** Right hippocampal 3D models for the same subject, constructed using all subfields from both methods in 3D Slicer[2]; boxplots represent mean hippocampal volumes for left and right hippocampi in HC expressed in mm3 and in percent of TIV. Upper models show anterior–superior view, and lower images represent inferior projections for comparison. Shaded gray-wireframe area embodies whole hippocampus representation created from standard FS segmentation; note reduced size of the vB model. Most noticeable subregion differences are related to the definition of the anterior and posterior extent of CA1 and posterior subiculum; more medially and dorsally extended in vB. Coincidentally, the hippocampal tail, pre-subiculum, and para-subiculum regions defined in FS represents at least partially overlapping areas between methods. Other deep internal hippocampal structures such as the molecular layer, GCMLDG, fissure, and fimbria are individually ascribed only in FS **(C)**. CA4-DG and CA2-CA3 subfields are jointly segmented in vB; CA4 and GCMLDF are grouped together in FS for comparison purposes. Volume differences are probably not only related to atlas definition; both approaches also show methodological discrepancies for intracranial volume computation. *Significant after Bonferroni correction. **Significant uncorrected $p < 0.05$. Paired-sample T-test for inter-hemispheric comparison of volumes as percent of TIV in HC.

show good test–retest reliability across scanner manufacturers and across field strengths (35, 36).

The FreeSurfer v6.0 algorithm follows a generative, parametric approach which focuses on modeling the spatial distribution of the hippocampal subregions and surrounding brain structures, which is learned from labeled training data. FreeSurfer v6.0 is built with a novel atlasing algorithm and *ex vivo* MRI data from autopsy brains. The segmentation provides 15 different subregions (12 used in for this work), based on the histology and morphometry from Rosene and Van Hoesen (37) and partly also on (38–41). See **Figure 1** for details on implemented atlas and segmentation.

The *ex vivo* imaging protocol yields images with high resolution and signal-to-noise ratio. The segmentation algorithm is similar to Van Leemput (42) which is appropriate for analyzing *in vivo* MRI scans of different manufacturers using different T1 contrasts.

Compared to other new methods available, FreeSurfer involves a prolonged processing time 8∼24 h running on standard single-core systems but also yielding extended quantification of additional brain structures including whole-brain regions beyond hippocampal formations.

We transformed the fixed-width-column plain-text files in which were written down the FreeSurfer outputs to comma separate values (csv) plain-text files which are more suitable to be opened as a Pandas' Dataframe (Python package). To ensure that classifiers did not consider features lacking specific regional information, we eliminated general features like cortical volume, mean cortical thickness, brain volume, and ventricle volume. Finally, to avoid potential biases due to differences among the participants' head size (43), volume measures of each area were normalized as a percentage of the estimated total intracranial volume (eTIV), provided also in FreeSurfer's results.

VolBrain—HIPS

VolBrain is a patch-based segmentation method for high-resolution hippocampus subfields. It has been validated and uses two publicly available segmentation protocols different from FreeSurfer on manually *ex vivo* segmented datasets (44, 45).

Both hippocampal segmentation protocols are available in volBrain-HIPS; Winterburn atlas disclosing 5 subregions was used for this work because it is more similar to the FreeSurfer v6.0 definition than Kulaga-Yoskovitz. VolBrain-HIPS is based on the combination of MOPAL (46), a multi-contrast extension

of the OPAL (47) patch-based label fusion segmentation method and a novel neural network-based error corrector. The method uses an adaptation of MOPAL, a patch matching segmentation method to produce fast and accurate T1 brain segmentations. The method also works on standard MRI acquisition with image resolution of clinical practice as well as on single T1w or single T2w images. The VolBrain approach performs well also on mono-contrast T1w and T2w images as well as under standard resolution images that are upsampled using the LASR (48, 49) super-resolution method. The HIPS method also includes an error corrector post-processing step based on the use of a boosted ensemble of a neural network algorithm that is proposed to minimize systematic segmentation errors at post-processing. It works in a fully automated manner providing accurate results outperforming state-of-the-art methods such as MAGeT (50), ASHS (51), and SurfPatch (52) which usually require extended periods of computing time. VolBrain-HIPS takes <20 min and performs fast segmentation as well as subject-specific library registration that only requires estimating one non-linear registration over small regions to translate the whole library to the case to be segmented.

Finally, an online report is generated and results are plotted as absolute or percent values adjusted for intracranial volume against a normal reference standard for each anatomical region. Segmentation images can also be downloaded for evaluation purposes.

The same T1 volumetric images used for FreeSurfer v6.0 were uploaded to VolBrain-HIPS[3] for this analysis, using Winterburn atlas definition for controls and patients (45). The produced final report including absolute values (mm3) and normalized to percent of brain volume were recorded for analysis.

SUPERVISED CLASSIFICATION WITH A RANDOM FOREST ALGORITHM

As suggested by several previous publications (53–59), the quantification of non-hippocampal volumes in HS patients usually shows widespread modifications, involving the thalamus, amygdala, subcortical temporal white matter, temporal pole, and entorhinal cortex among others.

To study structural changes in the brain without any bias, we used FreeSurfer v6.0 metrics, specifically parcels of cortical thickness and volumes of all the cerebral structures in combination with machine learning methods based on Random Forest Classifiers (RFC) (60). This process was based on the implementation of an automatic classification algorithm to evaluate group discrimination performance considering morphometric contribution of whole-brain structures as independent features, without any a priori consideration. The selection of RFC was based by several premises: (i) We were interested in considering linear and more importantly non-linear relationships between all the features. (ii) As the number of samples was relatively low (although it is high for this type of studies), the parameter tuning should be an optional step. (iii) The interpretability of the relevant features in the classification should be clear. Given these conditions and the experience of the research team, we selected RFC as the best suitable algorithm for the analysis (61–63).

Preprocessed features of cortical–subcortical volumes and cortical thickness normalized to estimated total intracranial volume (eTIV) were analyzed via a progressive feature elimination (PFE) procedure (64) with a Monte Carlo cross-validation scheme (33). Briefly, we performed 30 shuffle-splits in which we randomly selected 80% of the samples (with balanced classes) to train the RFC and the remaining 20% for testing to optimize the accuracy of RFC by varying the number of features from all to a single one according to its classificatory relevance. RFC quantifies a feature's importance depending on how much the average Gini impurity index decreases in the forest due to its use as a node in a tree (65). We used this score to progressively eliminate features by removing the feature with the lowest importance at each iteration. Finally, we kept the N first features in the ranking, where N is the optimal number of features such that using more than N features fails to improve the classifier's performance.

The optimal number of features was selected visually by indicating the minimal quantity at which accuracy became constant. We used this fixed number of features to compute the accuracy, the confusion matrix, and the ROC curve, and to obtain each subject's probability of being in each group (HC, left HS, and right HS).

We implemented this processing framework to perform three classifications: (i) a binary classification to discriminate HC and HS; (ii) a binary classification to discriminate left HS and right HS; and (iii) a multiclass classification to discriminate HC, left HS, and right HS. For each classification, we obtained the optimal number of features, the list of defined features, and the classification performance metrics (accuracy, confusion matrix, and ROC curves). Asymmetry metrics were not included in these analyses given the conceptual basis that RFCs consider the relationship between features, and therefore the asymmetry between hemispheres regions was indirectly taken into account.

These analyses were performed with the RFC implemented in the Python's scikit-learn package, with a fixed number of trees (2000) and the recommended number of features (P) in each split, where P is the square root of the full set of features. The maximum depth in each tree was not restricted a priori, i.e., nodes were expanded until all leaves were pure or until all leaves contained less than two samples.

Statistical Analysis

Results were analyzed independently for each method using the Statistical Package for Social Sciences SPSS (Version 23; IBM, Armonk, New York). Volume mean average and 95% confidence intervals (CI) were calculated for each hemisphere. Receiver operating characteristic analyses were used to obtain optimal sensitivity, specificity, and 95% CI computed for left and right hippocampal sclerosis patients. A normalization of absolute

[3]https://volbrain.upv.es/

values related to the total TIV was implemented and used for group comparison and correlation tests, since it was previously described as the most significant covariate to be considered (25). Normalization was performed by the following expression:

$$\text{normalized\%TIV}_{\text{Subject}} = \text{AbsoluteValue(cm3)}_{\text{Patients}}\text{X100/TIV}_{\text{Subject}}.$$

Both methods implement different atlas definitions and strategies to quantify TIV, thus precluding a direct comparison between absolute values.

Asymmetry degree was analyzed as an independent measure representing the difference between right and left regions divided by their mean (in percent) as implemented in vB and used in previous reports (25). Thus, positive values represent greater volumes on the right side.

Nominal variables were compared using the Chi square test. Paired-sample *t*-test (right vs. left) and ANCOVA (between groups) were used for normally distributed scalar variables adjusted for age and sex. Correlations were tested using the two-tailed Pearson coefficient controlling for age and sex. Significance level was adjusted for the effect of multiple comparisons using Bonferroni correction when appropriate. To test the difference between HS sides in the group of patients, an ANCOVA test was calculated on z-scores computed for each region using the following formula:

$$\text{z} - \text{score} = \frac{(\text{normalized\%TIV}_{\text{Patients}} - \text{normalizedmean\%TIV}_{\text{HC}})}{\text{standarddeviationmean \%TIV}_{\text{HC}}}.$$

Age, sex, and clinical characteristics of epilepsy were included in the analysis as covariates.

RESULTS

After correction for TIV, no significant correlation was found between age and sex with hippocampal or subfield volumes ($p > 0.05$) in controls or patients.

Controls and patients were paired according to age and sex, with female prevalence (controls 44f/17m, right HS 15f/7m, and left HS 17f/18m) not reaching significant differences (p.062). Groups were not different in relation to participants' age (p. 495), control subjects with a mean of 32 (18–62y), right HS patients group with 33 (21–64y), and left HS with 34 (19–52y).

Clinical characteristics of epilepsy including seizure frequency, age at onset, and epilepsy evolution time were similar ($p > 0.05$) in both groups of patients. Right HS patients had 7 (1–30) seizures per month, disease onset at 10 (1–40y), with a duration of 23 (6–40y), and left HS epilepsy patients presented 12 (1–90), 11 (1–32y), and 22 (2–49y), respectively.

No correlation was found between clinical features of epilepsy and hippocampus or subregion volumes.

Hippocampal Results

Estimated hippocampal volume and 95% confidence interval (CI) for controls on the right side were 3,454 cm3 (3.355–3.554)/0.2239% (0.2196–0.2283) for FS, and 2,480 cm3 (2.326–2.490)/0.1750% (0.1713–0.1787) for vB. Results for the left hippocampus were 3,398 cm3 (3.300–3.496)/0.2230% (0.2158–0.2248) for FS, and 2,320 cm3 (2.246–2.394)/0.1686% (0.1653–0.1720) for vB. Volume asymmetry was 1.6% (0.5–2.7) for FS and 3.6% (2.2–5) for vB with significant rightward lateralization ($p < 0.003$).

Mean ipsilateral hippocampal volume and 95% CI for right HS patients were 2,578 cm3 (2.401–2.755)/0.1743% (0.1603–0.1882) for FS and 1,429 cm3 (1.295–1.563)/0.1073% (0.0979–0.1167) for vB. Left hippocampus volume and 95% CI for the left HS patients were 2,560 cm3 (2.425–2.696)/0.1693% (0.1604–0.1783) for FS and 1,437 cm3 (1.324–1.549)/0.1055% (0.0981–0.1129) for vB.

Hippocampal asymmetry in the right HS group was −27.4% (−31.4/−23.5) for FS and −47% (−53.2/−41.6) for vB. Asymmetry in left HS patients was 33% (28.2/37.7) on FS and 53.2% (48–58.3) for vB. Hippocampal volumes ipsilateral to the HS side were significantly reduced compared with controls and also with the non-lesional side of right and left HS groups (p.000). Additionally, the right hippocampus was greater in left HS patients than in HC (FS, p.022) (see details in **Tables 1** and **2**).

Hippocampal asymmetry was the most reliable indicator for accurate classification between HC and right and left HS with an AUC:1 for Vb (measured in cm3 and in brain percent), an AUC:0.998 (using cm3), and an AUC:0.977 (in brain percent) based on FS. Optimal sensitivity–specificity was also calculated using hippocampal volumes with elevated accuracy (AUC:0.914 $\sim$ 0.993) for patient classification. Detailed results are specified in **Figure 2** and **Table 3**.

To specifically account for atrophy differences among HS sides, z-score volumes for each hippocampus were compared, and no significant differences were found (p.692, FS and p.768, vB).

Results for Hippocampal Subfields

The mean volume and 95% CI estimates of hippocampal subfields for HC and patients are detailed in **Tables 1** and **2**. In the HC group, a significant rightward asymmetry of hippocampal subfields was recognized for CA1, CA2–CA3, and CA4-DG (in vB) and for CA1, CA3, CA4, molecular layer, hippocampal fissure, and GC-ML of DG (in FS). Leftward lateralization was recognized for the subiculum (vB) and pre-subiculum (FS) subregions (see details in **Tables 1** and **2**).

All subregions on the ipsilateral side of HS patients showed significant volume reduction compared with HC using vB, and most subfields were also reduced considering FS except for the right (p.446) and left (p.140) HATA, right and left fissure (p.1), and right ipsilateral fimbria (p.849).

Most hippocampal subfields contralateral to the sclerotic side in left HS patients revealed greater volumes compared with HC, specifically right CA1(p.048 in vB), CA1(p.002 in FS) and CA3(p.016), CA4 (p.025), HATA (p.009), molecular layer, and GC-ML-DC (p.018) on FS. Only the left subiculum (p.035 vB) of right HS patients was reduced compared with HC.

The only subregion with a significant volume difference between sides of the affected hemisphere in patients was CA2–CA3 (p.024) for the group of right HS patients (observed in vB). Accordingly, ipsilateral to the sclerotic side, CA2–CA3 (vB) and CA3 (FS) subfields in left HS patients were less atrophic than any other cornus ammonia division.

TABLE 1 | volBrain-HIPS results.

Region	HC n:61		Right HS n:22		Left HS n:35	
	Mean & 95% CI volume (cm3)/ TIV-adjusted volume (%)	**Volume asymmetry percent (%)**	**Mean & 95% CI volume (cm3)/ TIV-adjusted volume (%)**	**Volume asymmetry percent (%)**	**Mean & 95% CI volume (cm3)/ TIV-adjusted volume (%)**	**Volume asymmetry percent (%)**
Right hippocampus	**2,408** (2.326–2.490) /**0.1750** (0.1713–0.1787)	**3.6** (2.2/5)**	**1,429** (1.295–1.563) /**0.1073** (0.0979–0.1167) § **2,293** (2.138–2.461) /**0.1727** (0.1614–0.1840)	**-47.4** (−53.2/−41.6)	**2,460** (2.331–2.590) /**0.1810** (0.1723–0.1896) **1,437** (1.324–1.549) /**0.1055** (0.0981–0.1129) Ψ	**53.2** (48/58.3)
Left hippocampus	**2,320** (2.246–2.394) /**0.1686** (0.1653–0.1720)					
Right CA1	**0.8084** (0.7772–0.8396) /**0.0587** (0.0571–0.0603)	**3** (0.8/5.3)*	**0.4776** (0.4262–0.5290) /**0.0359** (0.0320–0.0398) § **0.8016** (0.7400–0.8631) /**0.0603** (0.0555–0.0651)	**−51.6** (−58.5/−44.8)	**0.8524** (0.8009–0.9040) /**0.0626** (0.0592–0.06602) Ψ **0.4872** (0.4454–0.5290) /**0.0357** (0.0330–0.0384) Ψ	**55.1** (49.3/60.9)
Left CA1	**0.7834** (0.7541–0.8127) /**0.0569** (0.0554–0.0584)					
Right CA2–CA3	**0.1864** (0.1775–0.1953) /**0.0135** (0.0129–0.0141)	**21.7** (16.8/26.7)*	**0.0924** (0.0786–0.1063) /**0.0069** (0.0059−0.0079) § **0.1490** (0.1318–0.1663) /**0.0112** (0.0099–0.0124)	**−47.8** (−62.5/−33)	**0.1960** (0.1803–0.2117) /**0.0143** (0.0133–0.0154) **0.0821** (0.0727–0.0916) /**0.0060** (0.0053–0.0066) Ψ	**82.2** (73.9/90.4)
Left CA2–CA3	**0.1504** (0.1498–0.1589) /**0.0109** (0.0103–0.0114)					
Right CA4-DG	**0.6518** (0.6241–0.6796) /**0.0472** (0.0459–0.0486)	**8.1** (5.7/10)*	**0.3792** (0.3334–0.4250) /**0.0284** (0.0252–0.0317) § **0.5999** (0.5511–0.6488) /**0.0450** (0.0418–0.0481)	**−46.3** (−55.2/−37.4)	**0.6438** (0.6066–0.6810) /**0.0473** (0.0448–0.0499) **0.3566** (0.3193–0.3940) /**0.0261** (0.0236–0.0286) Ψ	**59.1** (51.7/66.6)
Left CA4-DG	**0.5996** (0.5764–0.6228) /**0.0435** (0.0424–0.0469)					
Right SR-SL-SM	**0.4828** (0.4649–0.5006) /**0.0350** (0.0342–0.0358)	**0.7** (−.9/2.5)	**0.2761** (0.2409–0.3113) /**0.0206** (0.0182–0.0230) § **0.4754** (0.4405–0.5104) /**0.0357** (0.0332–0.0382)	**−55** (-64.6/−45.4)	**0.5023** (0.4742–0.5303) /**0.0370** (0.0349–0.0390) **0.2819** (0.2544–0.3095) /**0.0207** (0.0188–0.0227) Ψ	**57.8** (51/54.5)
Left SR-SL-SM	**0.4780** (0.4623–0.4937) /**0.0347** (0.0339–0.0355)					
Right subiculum	**0.2792** (0.2686–0.2899) /**0.0203** (0.0196–0.0211)	**−9.9** (−12.4/-7.4)*	**0.2042** (0.1848–0.2237) /**0.0153** (0.0140–0.0165) § **0.2711** (0.2515–0.2907) /**0.0204** (0.0189–0.0218) §	**−28.6** (−35.8/−21.3)	**0.2662** (0.2506–0.2819) /**0.0196** (0.0184–0.0207) **0.2290** (0.2143–0.2438) /**0.0168** (0.0158–0.0179) Ψ	**15.3** (9.9/20.6)
Left subiculum	**0.3085** (0.2961–0.3201) /**0.0225** (0.0216–0.0233)					

*Paired-sample T-test; inter-hemispheric comparison in HC. *Significant after Bonferroni correction ($p < 0.001$). **Uncorrected ($p < 0.05$). Age- and sex-adjusted ANCOVA test between 3 groups; Bonferroni corrected ($p < 0.05$). Significant after pairwise comparisons: § Between controls and left HS. Ψ Between controls and right HS.*
Bold numbers are the mean values.

The most atrophic subfield ipsilateral to the sclerotic side for FS were CA4, GCMLDG, and molecular layer, and SLSRSM measured in vB in both right and left HS patients. See details in **Figure 3**.

Validation With the Automatic (Random Forest) Classifier

Our supervised machine learning validation process disclosed anatomical regions that were restricted to hippocampal subregions as the most relevant features to discriminate between patients and HC. In other terms, non-hippocampal regions were not identified as relevant for the classification.

The classifier was able to discriminate between controls and patients with a high accuracy in the three main classifications we performed: the classification between HC and patients (validation set mean accuracy: 0.907, AUC:0.960), between left and right HS patients (validation set mean accuracy: 0.91, AUC: 0.963), and between the three groups (validation set mean accuracy: 0.857, AUC: 0.960). The most important features and their relevance in each of three classifications are listed in **Figure 4**.

DISCUSSION

In this work, we define reference volumetric values and confidence intervals for hippocampus and hippocampal subfields using two commonly available approaches in a small community-based sample of healthy adults from Buenos Aires, Argentina. This is a limited sample but an important contribution to the field due to the scarce research literature on brain morphometric variations available in Latin America (66–68).

Since population variability on brain morphometric estimates are being increasingly reported (14, 69, 70, 70–72), it is important to consider the possibility of innate differences for adequate interpretation of MRI volumetry.

TABLE 2 | FreeSurfer v6.0 results.

Region	HC n:61		Right HS n:22		Left HS n:35	
	Mean & 95% CI volume (cm3)/TIV-adjusted volume (%)	**Volume asymmetry percent (%)**	**Mean & 95% CI volume (cm3)/TIV-adjusted volume (%)**	**Volume asymmetry percent (%)**	**Mean & 95% CI volume (cm3)/TIV-adjusted volume (%)**	**Volume asymmetry percent (%)**
Right hippocampus	**3,454** (3.355–3.554) /**0.2239** (0.2196–0.2283)	**1.6** (0.5/2.7)**	**2,578** (2.401–2.755) /**0.1743** (0.1603–0.1882)	**–27.4** (–31.4/–23.5)	**3,570** (3.404–3.737) /**0.2358** (0.2257–0.2459) Ψ	**33** (28.2/37.7)
Left hippocampus	**3,398** (3.300–3.496) /**0.2203** (0.2158–0.2248)		§ **3,386** (3204–3568) /**0.2289** (0.2134–0.2444)		**2,560** (2425–2696) /**0.1693** (0.1604–0.1783) Ψ	
Right CA1	**0.634** (0.611–0.657) /**0.0411** (0.0400–0.0421)	**3** (1.2/4.8)*	**0.483** (0.442–0.524) /**0.0326** (0.0296–0.0357)	**–27.6** (–33.2/–22)	**0.680** (0.645–0.715) /**0.0449** (0.0428–0.0471) Ψ	**37.8** (33.1/42.6)
Left CA1	**0.615** (0.594–0.636) /**0.0398** (0.0388–0.0408)		§ **0.636** (0.587–0.685) /**0.0428** (0.0397–0.0460)		**0.465** (0.435–0.494) /**0.0307** (0.0288–0.0326) Ψ	
Right CA3	**0.217** (0.208–0.225) /**0.0140** (0.0136–0.0145)	**13.1** (10.3/16)*	**0.157** (0.144–0.171) /**0.0106** (0.0097–0.0115)	**–21.6** (–28.2/–15)	**0.232** (0.218–0.246) /**0.0152** (0.0144–0.0161) Ψ	**43.4** (37.7/49.1)
Left CA3	**0.190** (0.183–0.197) /**0.0123** (0.0119–0.0127)		§ **0.194** (0.182–0.207) /**0.0132** (0.0121–0.0143)		**0.149** (0.139–0.159) /**0.0098** (0.0092–0.0104) Ψ	
Right CA4	**0.260** (0.252–0.268) /**0.0168** (0.0164–0.0172)	**5.9** (4/7.8)*	**0.182** (0.166–0.198) /**0.0123** (0.0111–0.0134)	**–31.5** (–37.4/–25.6)	**0.272** (0.258–0.287) /**0.0180** (0.0171–0.0188) Ψ	**43.9** (37.1/50.7)
Left CA4	**0.245** (0.237–0.252) /**0.0159** (0.0154–0.0163)		§ **0.248** (0.235–0.261) /**0.0168** (0.0155–0.0181)		**0.174** (0.162–0.186) /**0.0115** (0.0107–0.0123) Ψ	
Right presubiculum	**0.298** (0.288–0.307) /**0.0193** (0.0188–0.0198)	**–8.3** (–10.1/–6.5)*	**0.223** (0.207–0.239) /**0.0150** (0.0138–0.0163)	**–30.6** (–36.6/–24.7)	**0.293** (0.280–0.305) /**0.0194** (0.0185–0.0203)	**19** (13.8/24.1)
Left presubiculum	**0.324** (0.313–0.335) /**0.0210** (0.0204–0.0215)		§ **0.303** (0.285–0.320) /**0.0204** (0.0191–0.0218)		**0.243** (0.229–0.258) /**0.0161** (0.0151–0.0171) Ψ	
Right subiculum	**0.429** (0.416–0.443) /**0.0278** (0.0272–0.0285)	**–0.4** (–2.1/1.2)	**0.325** (0.298–0.351) /**0.0219** (0.0200–0.0237)	**–29.7** (–34.2/–25.3)	**0.434** (0.414–0.453) /**0.0286** (0.0274–0.0299)	**26.1** (21.4/30.8)
Left subiculum	**0.432** (0.417–0.446) /**0.0280** (0.0273–0.0286)		§ **0.436** (0.408–0.464) /**0.0294** (0.0273–0.0316)		**0.333** (0.316–0.350) /**0.0220** (0.0209–0.0232) Ψ	
Right parasubiculum	**0.58** (0.56–0.61) /**0.0038** (0.0036–0.0039)	**–3.9** (–7.6/–0.1)	**0.47** (0.43–0.51) /**0.0032** (0.0029–0.0035)	**–17.9** (–24.4/–11.4)	**0.58** (0.55–0.62) /**0.0039** (0.0036–0.0041)	**12.4** (5.1/19.8)
Left parasubiculum	**0.61** (0.58–0.63) /**0.0039** (0.0038–0.0041)		§ **0.56** (0.52–0.61) /**0.0038** (0.0035–0.0041)		**0.52** (0.47–0.57) /**0.0034** (0.0031–0.0037) §Ψ	
Right tail	**0.539** (0.519–0.558) /**0.0349** (0.0338–0.0360)	**–0.9** (–3.4/1.4)	**0.396** (0.366–0.426) /**0.0268** (0.0244–0.0292)	**–27.6** (–31.3/23.8)	**0.544** (0.513–0.575) /**0.0359** (0.0340–0.0378)	**30** (25.4/35.5)
Left tail	**0.543** (0.525–0.562) /**0.0353** (0.0342–0.0363)		§ **0.524** (0.483–0.565) /**0.0355** (0.0320–0.0390)		**0.399** (0.376–0.423) /**0.0265** (0.0248–0.0282) Ψ	
Right fissure	**0.148** (0.142–0.154) /**0.0096** (0.0092–0.0100)	**6** (2.6/9.3)*	**0.140** (0.126–0.154) /**0.0094** (0.0085–0.0103)	**–1.1** (–7.4/5.1)	**0.152** (0.142–0.161) /**0.0100** (0.0095–0.0105)	**8.9** (2/15.9)
Left fissure	**0.140** (0.133–0.146) /**0.0090** (0.0087–0.0094)		**0.142** (0.128–0.155) /**0.0096** (0.0086–0.0105)		**0.139** (0.129–0.148) /**0.0091** (0.0086–0.0097)	
Right molecular layer	**0.573** (0.555–0.591) /**0.0371** (0.0363–0.0380)	**2.2** (0.9/3.5)*	**0.424** (0.392–0.456) /**0.0286** (0.0263–0.0310)	**–28.4** (–33.1/–23.7)	**0.596** (0.567–0.625) /**0.0393** (0.0375–0.0411) Ψ	**34.8** (29.9/39.7)
Left molecular layer	**0.560** (0.543–0.577) /**0.0363** (0.0355–0.0372)		§ **0.562** (0.531–0.593) /**0.0380** (0.0354–0.0406)		**0.419** (0.395–0.444) /**0.0277** (0.0261–0.0293) Ψ	
Right GC-ML-DG	**0.302** (0.293–0.312) /**0.0196** (0.0191–0.0201)	**5.4** (3.7/7.1)*	**0.213** (0.195–0.231) /**0.0144** (0.0130–0.0158)	**–30.6** (–35.9/–25.2)	**0.318** (0.301–0.335) /**0.0210** (0.0199–0.0220) Ψ	**43** (36.6/49.5)
Left GC-ML-DG	**0.287** (0.277–0.296) /**0.0186** (0.0181–0.0191)		§ **0.288** (0.272–0.303) /**0.0195** (0.0180–0.0210)		**0.205** (0.192–0.219) /**0.0135** (0.0127–0.0144) Ψ	

(Continued)

TABLE 2 | Continued

Region	HC n:61		Right HS n:22		Left HS n:35	
	Mean & 95% CI volume (cm3)/TIV-adjusted volume (%)	**Volume asymmetry percent (%)**	**Mean & 95% CI volume (cm3)/TIV-adjusted volume (%)**	**Volume asymmetry percent (%)**	**Mean & 95% CI volume (cm3)/TIV-adjusted volume (%)**	**Volume asymmetry percent (%)**
Right fimbria	**0.81** (0.75–0.86) /**0.0052** (0.0049–0.0054)	**–0.2** (–5.2/4.6)	**0.71** (0.63–0.79) /**0.0048** (0.0041–0.0056)	**–7.2** (–15/.5)	**0.75** (0.69–0.81) /**0.0049** (0.0046–0.0053)	**14.2** (7.4/21)
Left fimbria	**0.81** (0.76–0.86) /**0.0052** (0.0049–0.0055)		**0.77** (0.67–0.86) /**0.0052** (0.0045–0.0059)		**0.65** (0.60–0.71) /**0.0043** (0.0039–0.0047) Ψ	
Right HATA	**0.59** (0.56–0.61) /**0.0038** (0.0036–0.0039)	**3.4** (0.3/6.5)	**0.52** (0.47–0.57) /**0.0035** (0.0031–0.0039) §§	**–8.4** (–17.2/.4)	**0.63** (0.60–0.66) /**0.0042** (0.0039–0.0044) Ψ	**23.8** (16.6/31.1)
Left HATA	**0.57** (0.54–0.59) /**0.0036** (0.0035–0.0038)		**0.57** (0.52–0.62) /**0.0038** (0.0035–0.0042)		**0.50** (0.47–0.53) /**0.0033** (0.0031–0.0035) ΨΨ	

*Paired sample T-test; inter-hemispheric comparison in HC. *Significant after Bonferroni correction, **Uncorrected $p < 0.05$. Age- and sex-adjusted ANCOVA test between 3 groups; Bonferroni corrected ($p < 0.05$). Significant after pairwise comparisons: § Between controls and left HS; Ψ between controls and right HS. §§ With left HS. ΨΨ With right HS. §Ψ Only with HC.*

Bold numbers are the mean values.

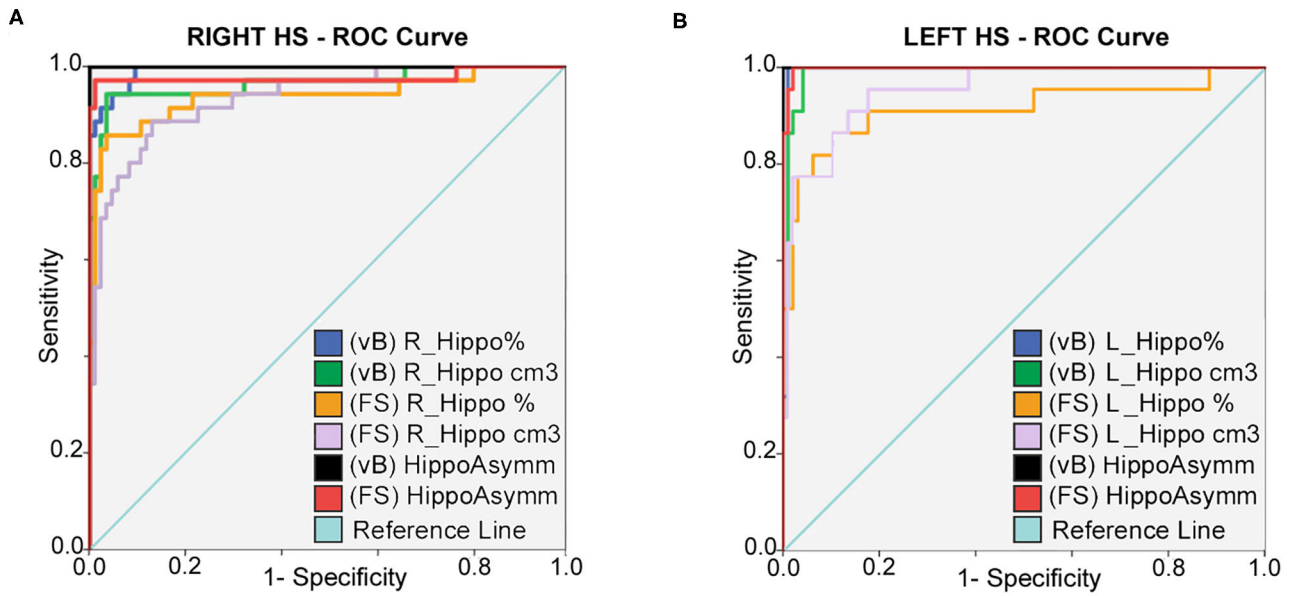


FIGURE 2 | ROC curves: sensitivity and specificity for classification between HC; right **(A)** and left **(B)** HS patients. Prediction is calculated for hippocampal volumes in percent of TIV; expressed in cm3 and also using asymmetry differences for both methods. Best accuracy was obtained using asymmetry values; hippocampal volume estimates from vB showed slightly better accuracy than FS.

Several methods provide quantification of brain structures by using MRI data, including freely available softwares and online processing services that usually report adjusted values considering intracranial total volume, age, and sex as covariates. Unfortunately, wide variability exists related to the employed methodology that impairs appropriate comparisons of results between different techniques. Results are usually matched against a mixture of publicly available database of normal subjects that may not entirely account for variation among populations. Thus, absence of local references for normal and pathologic hippocampus volumes may also be a challenge for non-neuroimaging experts.

In this work, we report volumes of hippocampal structures and subregions that are specific for two different methods, evaluating patients from Latin America. The proposed reference values are intended to clarify the results obtained using two different methodologies, which are based on unequal anatomical definitions, and therefore the resulting scores cannot be directly used for cross-comparisons (see details in **Figure 1**).

TABLE 3 | ROC curve: reference values for highest sensitivity and specificity detection for right **(A)** and left **(B)** HS; using hippocampal volumes and asymmetry values from both methods.

Right HS	Ref. value	Sens	Spec-1	AUC	Sig.
A					
R_Hippo (vB) %	0.1399	1.000	0.010	0.993	0.000
R_Hippo (vB) Cm3	1,910	0.909	0.031	0.989	0.000
R_Hippo (FS) %	0.1977	0.818	0.063	0.914	0.000
R_Hippo (FS) Cm3	3,044	0.864	0.104	0.952	0.000
Hippocampal (vB) Asym	−11.9	1.000	0.000	1.0	0.000
Hippocampal (FS) Asym	−8.45	1.000	0.21	0.998	0.000
B					
Left HS					
L_Hippo (vB) %	0.1442	0.943	0.048	0.992	0.000
L_Hippo (vB) Cm3	1,855	0.943	0.036	0.966	0.000
L_Hippo (FS) %	0.1920	0.857	0.036	0.937	0.000
L_Hippo (FS) Cm3	2,950	0.800	0.084	0.934	0.000
Hippocampal (vB) Asym	23.4	1.000	0.000	1.0	0.000
Hippocampal (FS) Asym	9.25	0.971	0.012	0.977	0.000

We calculated mean volumes, confidence intervals, and cutoff estimations to recognize a regional sample of patients with confirmed unilateral mesial sclerosis and temporal lobe epilepsy with high sensitivity and specificity. Hippocampal asymmetry degree was the most accurate measure for classification regardless of the volumetry method used, as previously reported by others (3, 23, 31).

Our results are coincident with previous reports supporting rightward asymmetry for whole hippocampal volume not only in HC but also present in other animal species (73).

Interestingly, as recently reported (74), some hippocampal subregion volumes in our study were leftward lateralized in HC including the subiculum and pre-subiculum, the former based on volBrain and the latter on FreeSurfer. This discrepancy probably represents similar findings observed in overlapping areas related to known differences in atlas definitions (75) (see **Figure 1**).

Contrary to previous findings (31), our results did not show any significant correlation between hippocampus volume and its subfields with clinical features of epilepsy.

Few studies had focused on assessing subregion atrophy differences between HS sides based on imaging data. We found specific volume reduction of CA2–CA3(vB) in right HS patients with partial preservation in left HS patients. Future investigation using adequate methodology and involving a greater number of participants may confirm our findings. A distinctive pattern of modifications can be expected from left and right HS which are not usually considered on histopathology research, probably supporting differences in functional abilities (76–80).

To our knowledge, only one published study directly addressed asymmetry differences between hippocampal subregions among left and right HS patients using FS v6.0 (81). The authors found reduced contralateral volumes to the side of HS for presubiculum, HATA, and TAIL subfields. Unfortunately, information about known constitutional asymmetries present in HC (74) that could influence the results as in our analysis is not usually considered.

Another recent study used an approach similar to ours (but based on manual segmentation) and found greater (rather than reduced) volume of left subiculum (contrary to our findings) in right HS participants (32). Additionally, the authors also showed significant reduction of ipsilateral CA1 subfield compared against any other subregion on the sclerotic side.

An interesting observation from our analysis is a trend to find larger volumes on mesial–temporal structures contralateral to the side of HS in patients compared with HC. Diverse hippocampal subfields and also the hippocampus (FS) in the right (non-lesional) hemisphere of left HS patients support this assumption showing significant greater volumes compared to the same regions in healthy controls (**Table 2**). We should stress that in clinical practice the interpretation of hippocampal volumetry alone may not adequately identify some confirmed cases (∼10%) with compatible clinical and paraclinical findings of HS which may only show subtle signal intensity changes on T2/FLAIR images (5, 82). Furthermore, it is important to note that a small group (∼20%) of confirmed temporal lobe epilepsy patients without abnormal MRI finding will be postoperatively classified as "Gliosis only" without hippocampal sclerosis based on histopathology (83), showing no evidence of neuronal loss nor hippocampal volume reduction.

Supplementary functional imaging examinations are useful for diagnosis in temporal lobe epilepsy with HS and unremarkable MRI findings that may preserve normal hippocampal volumes. Interictal FDG-PET (2-[18F]-fluoro-2-D-deoxyglucose positron emission tomography) is a relatively widely available neuroimaging modality with high sensitivity (∼80%) to disclose abnormal cortex hypo-metabolism in temporal lobe epilepsy (84, 85). Importantly, about 20% of patients with confirmed hippocampal sclerosis and normal MRI

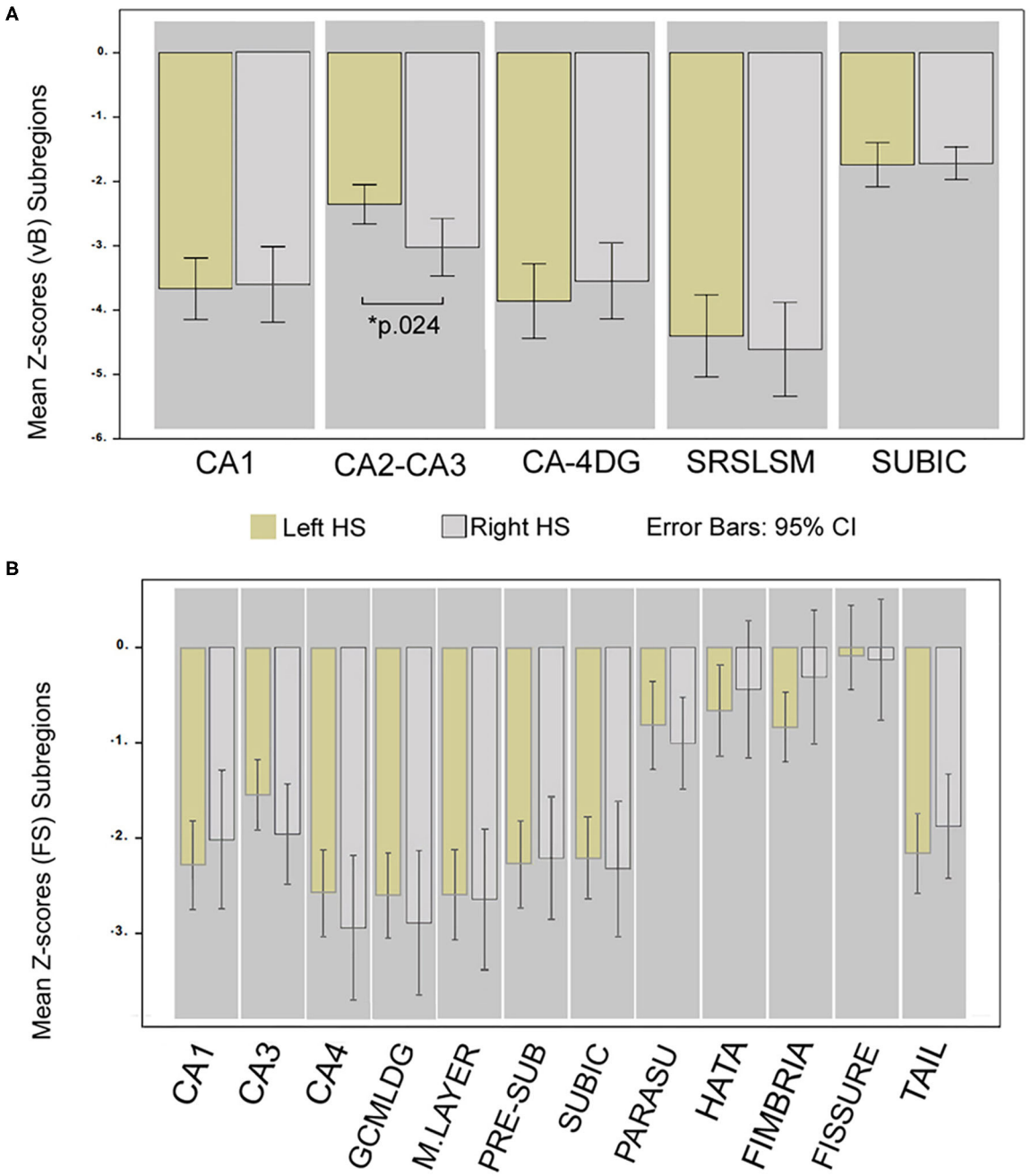


FIGURE 3 | Differences in hippocampal subregion atrophy; comparison of Z-scores between HS sides. Mean Z-score volume comparison between left and right HS; obtained from vB **(A)** and from FS **(B)**. *Significant for ANCOVA test between groups; Bonferroni corrected ($p < 0.05$) adjusted for age, sex, and epilepsy characteristics. Whiskers represent 95% confidence interval.

will show temporal cortex anomalies with reduced 18-FDG uptake (86, 87).

Although great progress has been made in recent years for preoperative diagnosis of HS using non-invasive methods, a considerable group of patients (20∼40%) will fail to achieve complete seizure free after surgery (88, 89) following appropriate medical practices in experienced epilepsy centers. A recognized limitation of our study is the absence of histopathology information about recent standardized ILAE classification for HS subtype (ILAE HS I-III) (90) that could allow us to correlate volumetry findings with specific subfield anomalies. Nevertheless, some controversies remain concerning the role of histopathologic classification for predicting clinical evolution in HS patients and also regarding the feasibility of MRI-histopathology correlations, limited by the amount of brain sample available for examination. Additional benefits of MRI

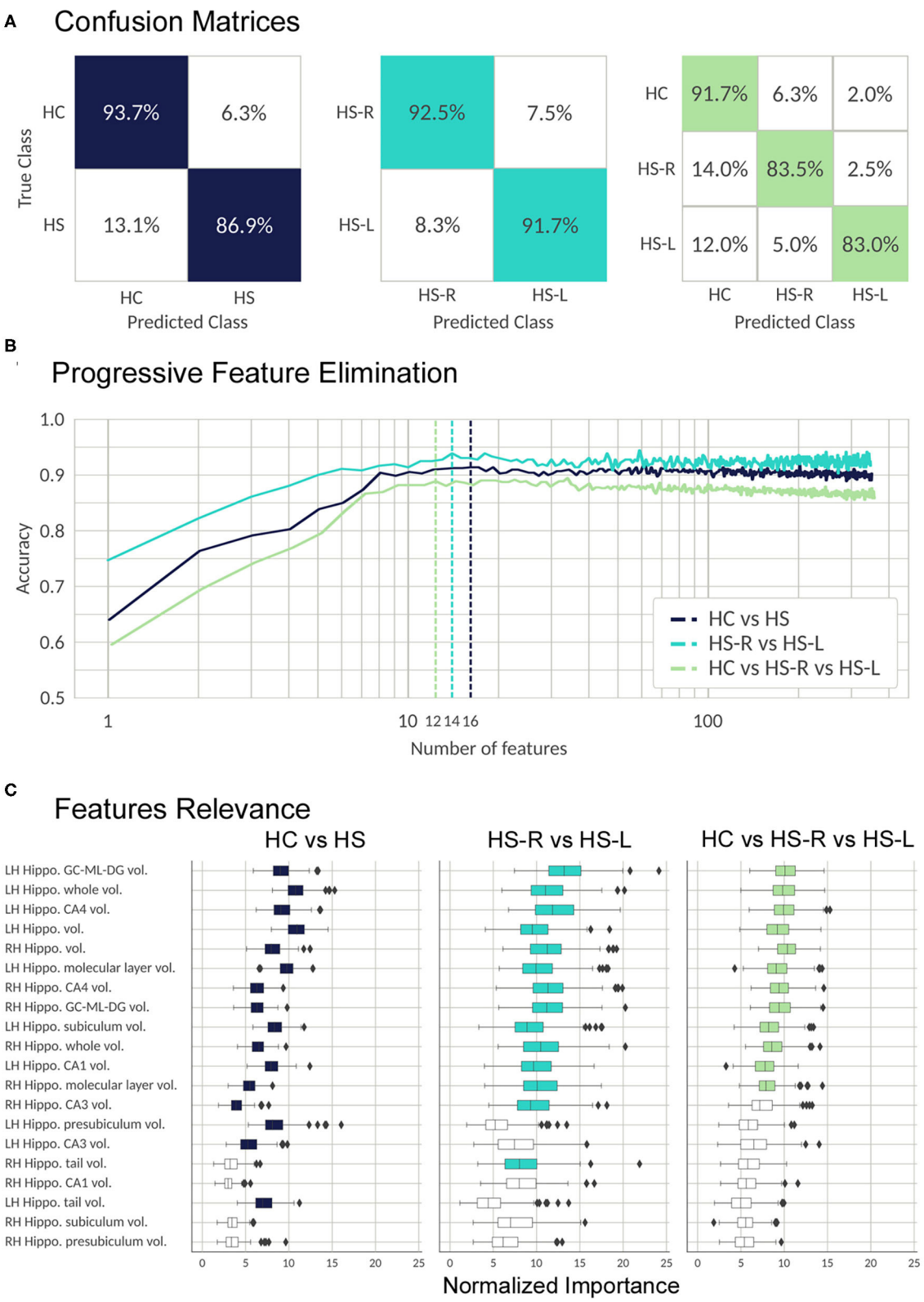


FIGURE 4 | Validation process; results from the RFC algorithm. **(A)** Random forest performance. The confusion matrices show the percentage of correct (colored) and incorrect predictions for each class. The value was accumulated over the Monte Carlo cross validation (200-folds). **(B)** Progressive feature elimination. Random forest

(Continued)

FIGURE 4 | mean accuracy over the Monte Carlo cross-validation (30-folds) as a function of the number of features used to train the model. The features are sorted from most to the least important. The dashed lines show the optimal number of features for each classifier. **(C)** Feature importance for the 20 most important FS metrics. The boxplots show the normalized random forest feature importance distribution over the Monte Carlo cross-validation (200-folds). The colored boxes are the features which were selected by the progressive feature elimination procedure. The feature importance value was normalized with respect to the trivial importance level 1/N, where N is the number of features—that means, at the trivial level all the features have the same importance. Whiskers represent 95% confidence interval; small rhombuses indicate outliers.

volumetry include the ability to examine the entire length of the sclerotic hippocampus and its contralateral homologous and also to consider inherited asymmetries for comparison.

Another caveat of this study is its relatively small sample size and also the uncertainty of segmentation accuracy of automated methods, to quantify structures on atrophic hippocampus. Some studies suggest that manual tracing methods may provide more accurate volumetric measurement than automated segmentation, especially in cases of HS (91, 92). However, validation results from FreeSurfer v6.0 developers indicate that subfield volumes still carry useful information, even when T1 images usually display limited contrast on the internal subregion boundaries (75). Equivalent methodology was also successfully implemented in previous studies on cognitive function and epilepsy (3, 4, 22, 23, 91, 93) with satisfactory results.

Contrary to previous observations supporting a fundamental role for cortical mesial–temporal regions, our machine learning-based validation process using an automatic algorithm failed to identify non-hippocampal structures such as the thalamus, temporal pole, fornix, or mammillary bodies as relevant for group classification. It shall be stressed that the abovementioned structures and others known to be involved in HS patients could falsely not been recognized as important due to a superior performance of hippocampal and subregion metrics in a trade-off between accuracy and number of analyzed features. Moreover, non-hippocampal anomalies preferentially involve white matter tracts (94, 95) and are usually related to prolonged epilepsy duration or high seizure frequency not considered in our validation process.

In conclusion, hippocampal anatomical structures are the most relevant features to recognize HS patients as confirmed by an automatic classification based on RFC. The local reference values proposed for hippocampal volumes and subfields may prove a useful guide for diagnosis in adult patients with temporal lobe epilepsy and suspected HS particularly for non-specialized radiologists.

Providing normal hippocampal reference values are a significant contribution to future studies focusing on regional morphometric variations in Latin America.

Finally, our results are also important for the interpretation of studies reporting hippocampal subfield volumes based on different atlas, which may show noticeable differences even when the same anatomical labels are used (96–98).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research and Bio-ethics review board at Hospital El Cruce, Carlos N. Kirschner. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

JP: study design, writing, statistical analysis, and quality revision. SK: study design and manuscript edition. PS: patients selection, clinical data, and follow-up. AN: patients selection, clinical data, and demography curation and analysis. SC: statistical analysis, data pre-processing, and quality revision. AD: data processing, analysis, classification algorithm, manuscript writing, and edition. GP: classification algorithm and manuscript writing and edition. JV: data pre-processing, analysis, quality revision, and manuscript edition. MV-A: data pre-processing, analysis, quality revision, and manuscript edition. PD-K: study design, data processing, analysis, classification algorithm, and manuscript writing and edition. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to acknowledge the kind contributions of colleagues from the Epilepsy Division, MRI-EEG technicians, and Administrative Personnel and especially from Prof. Paula Gonzalez at ENyS—El Cruce Hospital.

REFERENCES

1. Jack CR Jr, Sharbrough FW, Cascino GD, Hirschorn KA, O'Brien PC, Marsh WR. Magnetic resonance image-based hippocampal volumetry: correlation with outcome after temporal lobectomy. *Ann Neurol.* (1992) 31:138–46. doi: 10.1002/ana.410310204
2. Jack CR Jr, Petersen RC, O'Brien PC, Tangalos EG. MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology.* (1992) 42:183–8. doi: 10.1212/WNL.42.1.183
3. Louis S, Morita-Sherman M, Jones S, Vegh D, Bingaman W, Blumcke I, et al. Hippocampal sclerosis detection with neuroquant

compared with neuroradiologists. *AJNR Am J Neuroradiol.* (2020) 41:591–7. doi: 10.3174/ajnr.A6454
4. Vos S, Gavin P, Olivia G, Hugh G, Frederik B, Ferran P, et al. Hippocampal profiling: localized magnetic resonance imaging volumetry and T2 relaxometry for hippocampal sclerosis. *Epilepsia.* (2020) 61:297–309. doi: 10.1111/epi.16416
5. Bernasconi A, Cendes F, Theodore WH, Gill R, Koepp M, Hogan R, et al. Recommendations for the use of structural magnetic resonance imaging in the care of patients with epilepsy: a consensus report from the International League Against Epilepsy Neuroimaging Task Force. *Epilepsia.* (2019) 60:1054–68. doi: 10.1111/epi.15612
6. Lee DH, Gao FQ, Rogers JM, Gulka I, Mackenzie IR, Parrent AG, et al. MR in temporal lobe epilepsy: analysis with pathologic confirmation. *AJNR Am J Neuroradiol.* (1998) 19:19–27.
7. Cascino GD, Jack CR Jr, Parisi JE, Sharbrough FW, Hirschorn KA, Meyer FB, et al. Magnetic resonance imaging-based volume studies in temporal lobe epilepsy: pathological correlations. *Ann Neurol.* (1991) 30:31–6. doi: 10.1002/ana.410300107
8. Kim JH, Tien RD, Felsberg GJ, Osumi AK, Lee N, Friedman AH. Fast spin-echo MR in hippocampal sclerosis: correlation with pathology and surgery. *AJNR Am J Neuroradiol.* (1995) 16:627–36.
9. Fischl B. FreeSurfer. *NeuroImage.* (2012) 62:774–81. doi: 10.1016/j.neuroimage.2012.01.021
10. Romero JE, Coupé P, Manjón JV. HIPS: a new hippocampus subfield segmentation method. *Neuroimage.* (2017) 163:286–95. doi: 10.1016/j.neuroimage.2017.09.049
11. Bernhardt BC, Hong SJ, Bernasconi A, Bernasconi N. Magnetic resonance imaging pattern learning in temporal lobe epilepsy: classification and prognostics. *Ann Neurol.* (2015) 77:436–46. doi: 10.1002/ana.24341
12. Liang P, Shi L, Chen N, Luo Y, Wang X, Liu K, et al. Construction of brain atlases based on a multi-center MRI dataset of 2020 Chinese adults. *Sci Rep.* (2016) 5:18216. doi: 10.1038/srep18216
13. Isamah N, Faison W, Payne ME, MacFall J, Steffens D, Beyer J, et al. Variability in frontotemporal brain structure: the importance of recruitment of African Americans in neuroscience research. *PLoS ONE.* (2010) 5:e13642. doi: 10.1371/journal.pone.0013642
14. Pai PP, Mandal PK, Punjabi K, Shukla D, Gael A, Joon S et al. BRAHMA: population specific T1, T2, and FLAIR weighted brain templates and their impact in structural and functional imaging studies. *Magn Reson Imaging.* (2020) 70:5–21. doi: 10.1016/j.mri.2019.12.009
15. Mohandas AN, Bharath RD, Prathyusha PV, Gupta AK. Hippocampal volumetry: normative data in the Indian population. *Ann Indian Acad Neurol.* (2014) 17:267–71. doi: 10.4103/0972-2327.138482
16. Jalaluddin WM, Mat Jusoh N, Ali Basahai IA, Abdullah MS, Abdul Karim AH, Gazali AK. Normalised MRI volumetry of the hippocampus among normal malay children and adolescents. *Malays J Med Sci.* (2013) 20:31–8.
17. Embong MF, Yaacob R, Abdullah MS, Abdul Karim AH, Ghazali AK, Jalaluddin WM. MR volumetry of hippocampus in normal adult malay of age 50 years old and above. *Malays J Med Sci.* (2013) 20:25–31.
18. Gonçalves-Pereira PM, Oliveira E, Insausti R. Análisis volumétrico mesiotemporal: valores normativos del hipocampo, la amígdala y el córtex entorrinal en la población adulta portuguesa [Quantitative volumetric analysis of the hippocampus, amygdala and entorhinal cortex: normative database for the adult Portuguese population]. *Rev Neurol.* (2006) 42:713–22. doi: 10.33588/rn.4212.2005352
19. Gaillard WD, Bhatia S, Bookheimer SY, Fazilat S, Sato S, Theodore WH. FDG-PET and volumetric MRI in the evaluation of patients with partial epilepsy. *Neurology.* (1995) 45:123–6. doi: 10.1212/WNL.45.1.123
20. Szabo CA, Xiong J, Lancaster JL, Rainey L, Fox P. Amygdalar and hippocampal volumetry in control participants: differences regarding handedness. *AJNR Am J Neuroradiol.* (2001) 22:1342–5.
21. Li YJ, Ga SN, Huo Y, Li SY, Gao XG. Characteristics of hippocampal volumes in healthy Chinese from MRI. *Neurol Res.* (2007) 29:803–6. doi: 10.1179/016164107X223557
22. Sone D, Sato N, Maikusa N, Ota M, Sumida K, Yokoyama K, et al. Automated subfield volumetric analysis of hippocampus in temporal lobe epilepsy using high-resolution T2-weighed MR imaging. *Neuroimage Clin.* (2016) 12:57–64. doi: 10.1016/j.nicl.2016.06.008
23. Granados Sánchez AM, Orejuela Zapata JF. Diagnosis of mesial temporal sclerosis: sensitivity, specificity and predictive values of the quantitative analysis of magnetic resonance imaging. *Neuroradiol J.* (2018) 31:50–9. doi: 10.1177/1971400917731301
24. Nobis L, Manohar SG, Smith SM, Alfaro-Almagro F, Jenkinson M, Mackay C, et al. Hippocampal volume across age: nomograms derived from over 19,700 people in UK Biobank. *Neuroimage Clin.* (2019) 23:101904. doi: 10.1016/j.nicl.2019.101904
25. Santyr BG, Goubran M, Lau JC, Kwan B, Fateme S, Lee D, et al. Investigation of hippocampal substructures in focal temporal lobe epilepsy with and without hippocampal sclerosis at 7T. *J Magn Reson Imaging.* (2017) 45:1359–70. doi: 10.1002/jmri.25447
26. Pedraza O, Bowers D, Gilmore R. Asymmetry of the hippocampus and amygdala in MRI volumetric measurements of normal adults. *J Int Neuropsychol Soc.* (2004) 10:664–78. doi: 10.1017/S1355617704105080
27. Özdemir M, Soysal H, Eraslan Ö, Dilli A. Normative hippocampal volumetric measurements using magnetic resonance imaging. *Turk J Med Sci.* (2019) 49:1464–70.. doi: 10.3906/sag-1903-233
28. Laakso MP, Juottonen K, Partanen K, Vainio P, Soininen H. MRI volumetry of the hippocampus: the effect of slice thickness on volume formation. *Magn Reson Imaging.* (1997) 15:263–5. doi: 10.1016/S0730-725X(96)00390-6
29. Bhatia S, Bookheimer SY, Gaillard WD, Theodore WH. Measurement of whole temporal lobe and hippocampus for MR volumetry: normative data. *Neurology.* (1993) 43:2006–10. doi: 10.1212/WNL.43.10.2006
30. Farid N, Girard HM, Kemmotsu N, Smith M, Magda S, Lin W, et al. Temporal lobe epilepsy: quantitative MR volumetry in detection of hippocampal atrophy. *Radiology.* (2012) 264:542–50. doi: 10.1148/radiol.12112638
31. Kim JB, Suh SI, Kim JH. Volumetric and shape analysis of hippocampal subfields in unilateral mesial temporal lobe epilepsy with hippocampal atrophy. *Epilepsy Res.* (2015) 117:74–81. doi: 10.1016/j.eplepsyres.2015.09.004
32. Ji C, Zhu L, Chen C, Wang S, Zheng L, Li H. Volumetric changes in hippocampal subregions and memory performance in mesial temporal lobe epilepsy with hippocampal sclerosis. *Neurosci Bull.* (2018) 34:389–96. doi: 10.1007/s12264-017-0186-2
33. Xu QS, Liang YZ. Monte Carlo cross validation. *Chemom Intell Lab Syst.* (2001) 56:1–11. doi: 10.1016/S0169-7439(00)00122-2
34. Oddo S, Solis P, Consalvo D, Seaone E, Giagante B, Kochen S, et al. Postoperative neuropsychological outcome in patients with mesial temporal lobe epilepsy in Argentina. *Epilepsy Res Treat.* (2012) 2012:370351. doi: 10.1155/2012/370351
35. Han X, Jovicich J, Salat D, Avd Kouwe, Quinn BT, Czanner S et al. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage.* (2006) 32:180–94. doi: 10.1016/j.neuroimage.2006.02.051
36. Brown EM, Pierce ME, Clark DC, Fischl B, Iglesias J, Milberg W, 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
37. Rosene D, Van Hoesen, G. The hippocampal formation of the primate brain: a review of some comparative aspects of cytoarchitecture and connections. In: Jones E, Peters A. (Eds.), *Cerebral Cortex*. New York, NY: Plenum Press (1987). p. 6, 345–455. doi: 10.1007/978-1-4615-6616-8_9
38. Lorente de No, R. Studies on the structure of the cerebral cortex. II. Continuation of the study of the ammonic system. *J Psychol Neurol.* (1934) 46:113–77.
39. Insausti R, Amaral, D. *Hippocampal formation*. Mai J, Paxinos G. (Eds.) The Human Nervous System. (2011). p. 896–942. doi: 10.1016/B978-0-12-374236-0.10024-0
40. Green RC, Mesulam, M. Acetylcholinesterase fiber staining in the human hippo-campus and parahippocampal gyrus. *J Comp Neurol.* (1988) 273:488–99. doi: 10.1002/cne.902730405
41. Duvernoy, H. *The Human Hippocampus. An Atlas of Applied Anatomy*. J.F. Bergmann Verlag. Munich. (1988)
42. Van Leemput, K. Encoding probabilistic brain atlases using bayesian inference. *Med Imaging IEEE Trans.* (2009) 28:822–37. doi: 10.1109/TMI.2008.2010434

43. Whitwell JL, Crum WR, Watt HC, Fox NC. Normalization of cerebral volumes by use of intracranial volume: implications for longitudinal quantitative MR imaging. *AJNR Am J Neuroradiol.* (2001) 22:1483–9.
44. Yushkevich P, Avants B, Pluta J, Das S, Minkoff D, Mechanic-Hamilton D, et al. A high-resolution computational atlas of the human hippocampus from postmortem magnetic resonance imaging at 9.4 T. *NeuroImage.* (2009) 44:385–98. doi: 10.1016/j.neuroimage.2008.08.042
45. Winterburn JL, Pruessner JC, Chavez S, Schira M, Lobaugh N, Voineskos A, et al. A novel *in vivo* atlas of human hippocampal subfields using high-resolution 3 T magnetic resonance imaging. *Neuroimage.* (2013) 74:254–65. doi: 10.1016/j.neuroimage.2013.02.003
46. Romero JE, Coupe P, Manjón, J.V. *High Resolution Hippocampus Subfield Segmentation Using Multispectral Multi-atlas Patch-Based Label Fusion*. Patch-MI, LNCS *9993*. (2016) 117–24. doi: 10.1007/978-3-319-47118-1_15
47. Giraud R, Ta V.T, Papadakis N, Manjón J.V, Collins D.L, Coupé P. Alzheimer's Disease Neuroimaging Initiative. An optimized patchmatch for multi-scale and multi-feature label fusion. *NeuroImage.* (2016) 124:770–82. doi: 10.1016/j.neuroimage.2015.07.076
48. Manjón JV, Coupé P, Buades A, Fonov V, Collins DL, Robles, et al. Non-local MRI upsampling. *Med Image Anal.* (2010) 14:784–92.
49. Coupé P, Manjón JV, Chamberland M, Descoteaux M, Hiba, B. Collaborative patch- based super-resolution for diffusion-weighted images. *Neuroimage.* (2013) 83:245–61. doi: 10.1016/j.neuroimage.2013.06.030
50. Pipitone J, Park MTM, Winterburn J, Lett TA, Lerch JP, Pruessner C, et al. The Alzheimer's disease neuroimaging initiative, 2014 multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *NeuroImage.* (2014) 101:494–512. doi: 10.1016/j.neuroimage.2014.04.054
51. Yushkevich PA, Pluta JB, Wang H, Xie L, Ding SL, Gertje, et al. Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum Brain Mapp.* (2015) 36:258–87. doi: 10.1002/hbm.22627
52. Caldairou B, Bernhardt BC, Kulaga-Yoskovitz J, Kim H, Bernasconi N, Bernasconi, et al. *A Surface Patch-Based Segmentation Method For Hippocampal Subfields.* MICCAI, Part II, LNCS 9901. (2016) 379 - 387. doi: 10.1007/978-3-319-46723-8_44
53. Mahmoudi F, Elisevich K, Bagher-Ebadian H, Nazem-Zadeh M-R, Davoodi-Bojd E, Schwalb JM, et al. Data mining MR image features of select structures for lateralization of mesial temporal lobe epilepsy. *PLoS ONE.* (2018) 13:e0199137. doi: 10.1371/journal.pone.0199137
54. Mo J, Liu Z, Sun K, Ma Y, Hu W, Zhang C, et al. Automated detection of hippocampal sclerosis using clinically empirical and radiomics features. *Epilepsia.* (2019) 60:2519–29. doi: 10.1111/epi.16392
55. Rudie JD, Colby JB, Salamon N. Machine learning classification of mesial temporal sclerosis in epilepsy patients. *Epilepsy Res.* (2015) 117:63–9. doi: 10.1016/j.eplepsyres.2015.09.005
56. Cendes F, Andermann F, Gloor P, Evans A, Jones_Gotman M, Watson C, et al. MRI volumetric measurement of amygdala and hippocampus in temporal lobe epilepsy. *Neurology.* (1993) 43:719–25. doi: 10.1212/WNL.43.4.719
57. Bernhardt BC, Kim H, Bernasconi N. Patterns of subregional mesiotemporal disease progression in temporal lobe epilepsy. *Neurology.* (2013) 81:1840–7. doi: 10.1212/01.wnl.0000436069.20513.92
58. Garbelli R, Milesi G, Medici V, Villani F, Didoto G, Delea F, et al. Blurring in patients with temporal lobe epilepsy: clinical, high-field imaging and ultrastructural study. *Brain.* (2012) 135:2337–49. doi: 10.1093/brain/aws149
59. Kuzniecky R, Bilir E, Gilliam F, Faught E, Martin R, Hugg J. Quantitative MRI in temporal lobe epilepsy: evidence for fornix atrophy. *Neurology.* (1999) 53:496–501. doi: 10.1212/WNL.53.3.496
60. Donnelly-Kehoe PA, Pascariello GO, Gómez JC. Alzheimers Disease Neuroimaging Initiative. Looking for Alzheimer's Disease morphometric signatures using machine learning techniques. *J Neurosci Methods*. (2018) 302:24–34. doi: 10.1016/j.jneumeth.2017.11.013
61. Donnelly-Kehoe PA, Pascariello GO, García AM, Hodges J, Miller B, Rosen H, et al. Robust automated computational approach for classifying frontotemporal neurodegeneration: multimodal/multicenter neuroimaging. *Alzheimers Dement.* (2019) 11:588–98. doi: 10.1016/j.dadm.2019.06.002
62. Misra S, Wu Y. Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. *Machine Learn Subsurface Characterization*. (2019) 289–314. doi: 10.1016/B978-0-12-817736-5.00010-7
63. Palczewska A, Palczewski J, Robinson RM, & Neagu D. Interpreting random forest models using a feature contribution method. In I*EEE 14th International Conference on Information Reuse & Integration (IRI)*. San Francisco, CA: IEEE (2013). p. 112–9. doi: 10.1109/IRI.2013.6642461
64. Breiman, L. *Random forests. Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
65. Finkelsztejn A, Fragoso YD, Bastos EA, Duarte J, Santos Varela J, Houbrechts, et al. Intercontinental validation of brain volume measurements using MSmetrix. *Neuroradiol J.* (2018) 31:147–9. doi: 10.1177/1971400917721521
66. Curiati PK, Tamashiro JH, Squarzoni P, Duran F, Santos L, Wojngarten M, et al. Brain structural variability due to aging and gender in cognitively healthy Elders: results from the Sáo Paulo Ageing and Health study. *AJNR Am J Neuroradiol.* (2009) 30:1850–6. doi: 10.3174/ajnr.A1727
67. Squarzoni P, Tamashiro-Duran J, Souza Duran FL, Santos L, Vallada H, Menezes P, et al. Relationship between regional brain volumes and cognitive performance in the healthy aging: an MRI study using voxel-based morphometry. *J Alzheimers Dis.* (2012) 31:45–58. doi: 10.3233/JAD-2012-111124
68. Tang Y, Hojatkashani C, Dinov I, Sun B, Lingshong F, Lin XW, et al. The construction of a Chinese MRI brain atlas: a morphometric comparison study between Chinese and Caucasian cohorts. *Neuroimage*. (2010) 51:33–41. doi: 10.1016/j.neuroimage.2010.01.111
69. Rao NP, Jeelani H, Achalia R, Achiala G, Arpitha J, Bharat R, et al. Population differences in brain morphology: need for population specific brain template. *Psychiatry Res Neuroimaging.* (2017) 265:1–8. doi: 10.1016/j.pscychresns.2017.03.018
70. Yang G, Zhou S, Bozek J, Dong HM, Han M, Zuo XN, et al. Sample sizes and population differences in brain template construction. *Neuroimage.* (2020) 206:116318. doi: 10.1016/j.neuroimage.2019.116318
71. Sivaswamy J, Thottupattu AJ, Mehta R, Sheelakumari R, Kesavadas C. Construction of Indian human brain atlas. *Neurol India.* (2019) 67:229–34. doi: 10.4103/0028-3886.253639
72. Lee JS, Lee DS, Kim J, Kyeong K, Kong E, Kong H, et al. Development of Korean standard brain templates. *J Korean Med Sci.* (2005) 20:483–8. doi: 10.3346/jkms.2005.20.3.483
73. Freeman H, Cantalupo C and Hopkins W. Asymmetries in the hippocampus and amygdala of chimpanzees (Pan troglodytes). *Behav Neurosci.* (2004) 118:1460–5. doi: 10.1037/0735-7044.118.6.1460
74. Sarica A, Vasta R, Novellino F, Vaccaro M, Cerasa A, Quattrone A, et al. MRI asymmetry index of hippocampal subfields increases through the continuum from the mild cognitive impairment to the Alzheimer's disease. *Front Neurosci.* (2018) 12:576. doi: 10.3389/fnins.2018.00576
75. Iglesias JE, Augustinack JC, Nguyen K, Player CM, Player A, Wright M et al. A computational atlas of the hippocampal formation using *ex vivo*, ultra-high resolution MRI: application to adaptive segmentation of *in vivo* MRI. *Neuroimage.* (2015) 115:117–37. doi: 10.1016/j.neuroimage.2015.04.042
76. Coras R, Pauli E, Li J, Schwarz M, Rossler K, Buchfelder M, et al. Differential influence of hippocampal subfields to memory formation: insights from patients with temporal lobe epilepsy. *Brain.* (2014) 137:1945–57. doi: 10.1093/brain/awu100
77. Comper SM, Jardim AP, Corso JT, Botelho-Gaca L, Silva-Noffs M, Lancellotti C, et al. Impact of hippocampal subfield histopathology in episodic memory impairment in mesial temporal lobe epilepsy and hippocampal sclerosis. *Epilepsy Behav*. (2017) 75:183–9. doi: 10.1016/j.yebeh.2017.08.013
78. Daumas S, Halley H, Francés B, Lassalle JM. Encoding, consolidation, and retrieval of contextual memory: differential involvement of dorsal CA3 and CA1 hippocampal subregions. *Learn Mem.* (2005) 12:375–82. doi: 10.1101/lm.81905
79. Kesner RP, Lee I, Gilbert, P. A behavioral assessment of hippocampal function based on a subregional analysis. *Rev Neurosci.* (2004) 15:333–51. doi: 10.1515/REVNEURO.2004.15.5.333
80. Long J, Feng Y, Liao H, Zhou Q, Urbin MA. Motor sequence learning is associated with hippocampal subfield volume in humans

with medial temporal lobe epilepsy. *Front Hum Neurosci.* (2018) 12:367. doi: 10.3389/fnhum.2018.00367
81. Kreilkamp BAK, Weber B, Elkommos SB, Richardson MP, Keller SS. Hippocampal subfield segmentation in temporal lobe epilepsy: relation to outcomes. *Acta Neurol Scand.* (2018) 137:598–608. doi: 10.1111/ane.12926
82. Briellmann RS, Kalnins RM, Berkovic SF, Jackson GD. Hippocampal pathology in refractory temporal lobe epilepsy: T2-weighted signal change reflects dentate gliosis. *Neurology.* (2002) 58:265–71. doi: 10.1212/WNL.58.2.265
83. Blumcke I, Pauli E, Clusmann H, Schramm J, Becker A, Elger C et al. A new clinico-pathological classification system for mesial temporal sclerosis. *Acta Neuropathol.* (2007):113:235–44. doi: 10.1007/s00401-006-0187-0
84. Salanova V, Markand O, Worth R, Smith R, Wellman H, Hutchins, et al. FDG-PET and MRI in temporal lobe epilepsy: relationship to febrile seizures, hippocampal sclerosis and outcome. *Acta Neurol Scand.* (1998) 97:146–53. doi: 10.1111/j.1600-0404.1998.tb00628.x
85. Guedj E, Bonini F, Gavaret M, Trebuchon A, Aubert S, Boucekine M, et al. 18FDG-PET in different subtypes of temporal lobe epilepsy: SEEG validation and predictive value. *Epilepsia*. (2015) 56:414–21. doi: 10.1111/epi.12917
86. Henry TR, Roman DD. Presurgical epilepsy localization with interictal cerebral dysfunction. *Epilepsy Behav*. (2011) 20:194–208. doi: 10.1016/j.yebeh.2010.12.008
87. YangPF, Pei JS, Zhang HJ, Lin Q, Mei Z, Zhong Z-H, et al. Long-term epilepsy surgery out- comes in patients with PET-positive, MRI-negative temporal lobe epilepsy. *Epilepsy Behav*. (2014) 41:91–7. doi: 10.1016/j.yebeh.2014.09.054
88. LoPinto-Khoury C, Sperling MR, Skidmore C, Nei M, Evans J, Sharan A, et al. Surgical out- come in PET-positive, MRI-negative patients with temporal lobe epilepsy. *Epilepsia.* (2012) 53:342–48. doi: 10.1111/j.1528-1167.2011.03359.x
89. McIntosh AM, Wilson SJ, Berkovic SF. Seizure outcome after temporal lobectomy: current research practice and findings. *Epilepsia.* (2001) 42:1288–307. doi: 10.1046/j.1528-1157.2001.02001.x
90. Blümcke I, Thom M, Aronica E, Armstrong D, Bartolomei F, Bernasconi A, et al. International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: a task force report from the ILAE Commission on Diagnostic Methods. *Epilepsia.* (2013) 54:1315–29. doi: 10.1111/epi.12220
91. Steve TA, Gargula J, Misaghi E, Nowacki T, Schmitt L, Wheatley B, et al. Hippocampal subfield measurement and ILAE hippocampal sclerosis subtype classification with *in vivo* 4.7 tesla MRI. *Epilepsy Res.* (2020) 161:106279. doi: 10.1016/j.eplepsyres.2020.106279
92. Pardoe HR, Pell GS, Abbott DF, Jackson, G.D. Hippocampal volume assessment in temporal lobe epilepsy: how good is automated segmentation? *Epilepsia.* (2009) 50:2586–92. doi: 10.1111/j.1528-1167.2009.02243.x
93. Winston GP, Cardoso MJ, Williams EJ, Brudett J, Bartlett P, Espak M, et al. Automated hippocampal segmentation in patients with epilepsy: available free online. *Epilepsia*. (2013) 54:2166–73. doi: 10.1111/epi.12408
94. Besson P, Dinkelacker V, Valabregue R, Thivard L, Leclerc X, Baulac M, et al. Structural connectivity differences in left and right temporal lobe epilepsy. *Neuroimage*. (2014) 100:135–44. doi: 10.1016/j.neuroimage.2014.04.071
95. Rodríguez-Cruces R, Concha L. White matter in temporal lobe epilepsy: clinico-pathological correlates of water diffusion abnormalities. *Quant Imaging Med Surg.* (2015) 5:264–78.
96. Zammit AR, Ezzati A, Zimmerman ME, Lipton RB, Lipton ML, Katz MJ. Roles of hippocampal subfields in verbal and visual episodic memory. *Behav Brain Res*. (2017) 317:157–62. doi: 10.1016/j.bbr.2016.09.038
97. Ono SE, de Carvalho Neto A, Joaquim MJM, Dos Santos GR, de Paola L, Silvado CES. Mesial temporal lobe epilepsy: revisiting the relation of hippocampal volumetry with memory deficits. *Epilepsy Behav.* (2019) 100:106516. doi: 10.1016/j.yebeh.2019.106516
98. Peixoto-Santos JE, Carvalho LED, Kandratavicius L, Diniz PRB, Scandiuzzi RC, Coras R et al. Manual hippocampal subfield segmentation using high-field MRI: impact of different subfields in hippocampal volume loss of temporal lobe epilepsy patients. *Front. Neurol.* (2018) 9:927. doi: 10.3389/fneur.2018.00927

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 22 March 2021
doi: 10.3389/fnins.2021.625473

Diffusion Tensor Imaging Group Analysis Using Tract Profiling and Directional Statistics

Mehmet Özer Metin and Didem Gökçay*

Department of Health Informatics, Middle East Technical University, Ankara, Turkey

Group analysis in diffusion tensor imaging is challenging. Comparisons of tensor morphology across groups have typically been performed on scalar measures of diffusivity, such as fractional anisotropy (FA), disregarding the complex three-dimensional morphologies of diffusion tensors. Scalar measures consider only the magnitude of the diffusion but not directions. In the present study, we have introduced a new approach based on directional statistics to use directional information of diffusion tensors in statistical group analysis based on Bingham distribution. We have investigated different directional statistical models to find the best fit. During the experiments, we confirmed that carrying out directional statistical analysis along the tract is much more effective than voxel- or skeleton-guided directional statistics. Hence, we propose a new method called tract profiling and directional statistics (TPDS) applicable to fiber bundles. As a case study, the method has been applied to identify connectivity differences of patients with major depressive disorder. The results obtained with the directional statistic-based analysis are consistent with those of NBS, but additionally, we found significant changes in the right hemisphere striatum, ACC, and prefrontal, parietal, temporal, and occipital connections as well as left hemispheric differences in the limbic areas such as the thalamus, amygdala, and hippocampus. The results are also evaluated with respect to fiber lengths. Comparison with the output of the network-based statistical toolbox indicated that the benefit of the proposed method becomes much more distinctive as the tract length increases. The likelihood of finding clusters of voxels that differ in long tracts is higher in TPDS, while that relationship is not clearly established in NBS.

Keywords: diffusion tensor imaging, directional statistic, group analysis, tract profile, major depression

OPEN ACCESS

Edited by:
Tim B. Dyrby, Technical University of Denmark, Denmark

Reviewed by:
Longchuan Li, Marcus Autism Center, Children's Healthcare of Atlanta, United States
James Olav Breen Norris, Copenhagen University Hospital, Denmark

****Correspondence:***
Mehmet Özer Metin
ozermetin@gmail.com

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 November 2020*
Accepted: *12 February 2021*
Published: *22 March 2021*

Citation:
Metin MÖ and Gökçay D (2021) Diffusion Tensor Imaging Group Analysis Using Tract Profiling and Directional Statistics. Front. Neurosci. 15:625473. doi: 10.3389/fnins.2021.625473

INTRODUCTION

Diffusion tensor imaging (DTI) can reveal complicated structural differences in patient groups by using the orientation and integrity of white matter tracts to identify white matter abnormalities. The diffusion tensor is the covariance matrix of diffusion coefficients calculated from gradient directions for each voxel. Although DTI is by nature a nonscalar image which provides directional information for the neural tracts, group-based DTI analyses are mainly conducted using scalar descriptors such as fractional anisotropy (FA) (Basser, 1995), relative anisotropy (RA) (Basser and Pierpaoli, 2011),

axial diffusivity (AD), and radial diffusivity (RD) (Song et al., 2002). Such scalar metrics do not describe the full tensor shape or distribution and do not capture all of the information available in the data. By developing advanced metrics for connectivity analysis between groups of subjects in a nonscalar fashion, findings regarding abnormalities can be improved.

The principal diffusion direction (PDD), which is the eigenvector that corresponds to the largest eigenvalue of the tensor, captures the estimation of the fiber direction within the voxel. PDD has been used mainly in directionally encoded color (DEC) maps (Pajevic and Pierpaoli, 1999) which facilitate visual comparison but not quantitative group analysis. In order to evaluate PDD, which is a vector, statistical methods that analyze vector and tensor data are needed.

Directional statistics is conducted on vectors and directions based on observations on compact Riemannian manifolds (Pennec, 2006). Hence, it can encapsulate much more information than scalar metrics about the diffusion. Without the limitation of scalar statistics, one can evaluate dispersion and coherence values among the populations, fit directional model to the data, and perform hypothesis testing for group-based studies.

In the literature, directional statistics have been used to characterize fiber orientation distribution functions, to estimate fiber dispersion quantitatively *via* fanning and bending fiber geometries throughout the brain (Sotiropoulos et al., 2012; Tariq et al., 2016). In addition, directional statistics have also been utilized to extract bundle-specific metrics from crossing fiber models (Riffert et al., 2014) and fiber tractography (Parker et al., 2003). However, Watson distribution, which has been used in previous directional statistics in group analysis, contains limited parameters (Schwartzman et al., 2005; Hutchinson et al., 2012). Watson distribution is a bimodal probability distribution on a two-dimensional unit sphere S2 in R3 which is symmetrical around mean direction, where each direction and its negative have the same probability. In our previous study (Metin and Gökçay, 2014), it has been shown that Bingham distribution better fits into PDD distributions for white matter tracts and improves the depiction of variability among subjects in anisotropic tensors areas, such as fiber crossings. This is because Bingham distribution is a generalization of Watson distribution: it is bimodal and elliptic around mean direction.

Group analysis methods on DTI or DWI data can be classified into three: (1) region of interest (ROI)-based methods, (2) voxel-based analysis, and (3) fiber tract-based analysis. ROI-based methods are very labor intensive plus error-prone. On the other hand, voxel-wise comparison is open to misalignment of voxels because during registration of individual subject's data to a common space, topological variabilities may not be thoroughly resolved (Jones and Cercignani, 2010) for each fine structure. The amount of smoothing can greatly affect the final results, but there is no principled way of deciding how much smoothing is "correct" (Jones et al., 2005). For instance, tract-based spatial statistics (TBSS) tackles the alignment and smoothing problem for voxel-wise statistics by combining strengths of VBM-style analyses and tractography-based approaches (Smith et al., 2006). In short, analyses that involve fiber tracts are contingent upon computation of quantitative parameters of interest along the tracts (Goodlett et al., 2009) within diffusion tensor images. The properties of the fiber tract can be scalar values derived from tensors such as MD, FA, or trace, as well as shape information such as curvature and torsion of the specific tract (Mandl et al., 2010).

In this study, we propose a new tract-based framework using directional information in diffusion tensors to improve statistical group analysis, named as track profiling and directional statistics (TPDS). For this purpose, we have (1) generated a new data structure called tract profile by clustering fibers across subjects and (2) developed a method based on directional statistics to compare white matter (WM) differences of different groups across each tract profile. Overall, this new DTI group analysis method is called TPDS.

In order to demonstrate the superiority of the proposed framework, we compared the tract profiling method with two widely used techniques: TBSS (Smith et al., 2006) and voxel-based analysis (VBA) (Hecke et al., 2009). Furthermore, we ran a third comparison with the network-based statistic (NBS) toolbox (Zalesky et al., 2010) which utilizes nonparametric statistical testing to identify the components of an $N \times N$ undirected connectivity matrix that differ significantly between two distinct populations.

As a proof of concept, we demonstrated the strength of TPDS in the identification of differences of structural connectivity in major depressive disorder in a small data set (n = 30). Although depression has traditionally been viewed as an affective disorder, the last few decades of research have shown that MDD is also associated with considerable disturbances in cognitive functioning, including executive functions, attention, memory, and psychomotor speed (Castaneda et al., 2008; McClintock et al., 2010). In MDD, multidimensional, systems-level differences are reported in discrete, but functionally integrated pathways (Mayberg, 2003). Therefore, differences in MDD can be expected to cover a wide range of WM tracts. So far, especially white matter disturbances and connectivity differences have been analyzed using DTI-based analysis in MDD (Seminowicz et al., 2004; Zou et al., 2008; Cullen et al., 2010; Kieseppä et al., 2010; McClintock et al., 2010; Helm et al., 2018). Most of these studies state that loss of integrity occurs in the WM fiber tracts of the frontal, temporal, and cingulate cortex of MDD patients. White matter integrity can be described as biophysical white matter changes as a result of microstructural characteristic in both intra- and extra-axonal environments of WM such as axonal water fraction (AWF), intra-axonal diffusivity, and extra-axonal axial and radial diffusivities. More specifically, reported abnormalities in the connectivity of the DLPFC and ACC circuits (Helm et al., 2018), as well as subcortical regions, complement other findings specified in affective disorders (Sexton et al., 2009).

MATERIALS AND METHODS

Data Acquisition

In order to demonstrate the benefits of TPDS, we used T1-weighted, T2-weighted, and DTI MR data obtained from healthy subjects and patients with MDD.

Subjects

The control group consisted of 14 healthy subjects (8 female and 6 male) with age 31.71 ± 7.62, who had no history of neurological disease and also are not taking any medication. The depression group consisted of medication-naïve 16 subjects (8 female and 8 male, age: 31.12 ± 8.95)[1]. The data was collected as part of a local institutional project funded by METU (BAP-07-04-2012). Project management and subject recruitment were handled by a larger project[2] for which the results will be published elsewhere.

MRI Parameters

Whole-brain MRI scans were collected using the Siemens MAGNETOM 3 T scanner situated at the Bilkent University UMRAM center. T1-weighted [repetition (TR): 2,500 ms, echo time (TE): 3 ms, inversion time (TI): 1,000 ms, flip angle (FA): 8°, sagittal plane 1 mm isotropic resolution], T2-weighted (TR: 5,900 ms, TE: 108 ms, FA: 120°, spacing: 2.2, slice thickness 2 mm), and DWII scans (TR: 8,270 ms, TE: 83 ms, FA: 90°, spacing: 2.2, seven images with b-factor = 0 s/mm^2, 45 directions b-factor = 700 s/mm^2) are collected from the participants in a single session.

Data Processing

Pre Processing

We have implemented a fully automated pipeline to perform preprocessing as illustrated in **Figure 1**. The overall pipeline has been designed using the Connectome Mapper (Daducci et al., 2012). At the individual subject level, preprocessing steps are performed using several software toolkits. The first step is intrasubject registration of T1, T2, and DWI images using FSL's FLIRT as described in Jenkinson and Smith (2001) and Jenkinson et al. (2012). The registration is first done between the T2-weighted image and DWI B0 images, and then the high-resolution T1-weighted image is registered to the T2-weighted image. To eliminate the problem of transforming diffusion tensors, all of the images are registered to the DWI B0 image. This way, all image operations are performed on the diffusion image. For segmentation and parcellation of ROIs, FreeSurfer (Fischl et al., 2002) has been used. These steps transform the subject's MRI to uniform space and segment white and gray matter as well as cortical and subcortical structures based on the underlying atlas. The parcellation algorithm (Fischl et al., 2004) reveals 83 distinct cortical and subcortical structures of the brain using the Desikan–Killiany atlas (Desikan et al., 2006). All of these steps constitute the top row of **Figure 1**.

DTI processing begins with motion and eddy current artifact correction in FSL. Tensor estimation is done by Diffusion Toolkit (DTK) (Wang et al., 2007). For tractography (Parker et al., 2003; Cook et al., 2005), streamline fiber-tracking algorithm in Camino has been used. Each voxel in the parcellated image is selected as seeds. Eighty-three distinct cortical and subcortical areas are masked, and the generated binary image is used as the seed file of the algorithm for particular ROIs. For tracking, the fourth-order Runge–Kutta method has been chosen to propagate the tracks using a constant step size. Nearest-neighbor interpolation is applied around local voxel data. A minimum length criterion, 10 mm, is enforced to eliminate premature tract termination due to low SNR and low pathway anisotropy (Behrman-Lay et al., 2015). Each fiber bundle is pruned so that it only contains fibers connecting relevant regions. The number of streamlines depends on the size of the ROI. No additional elimination technique has been applied other than minimum length. These steps are illustrated in the second row of **Figure 1**.

Using the Connectome Mapper (Daducci et al., 2012), a connection matrix is generated to calculate the connectivity of the areas *via* the fiber tracts obtained in the first and second rows of **Figure 1**. After this step, the fiber tracts that connect corresponding brain areas will be bundled to construct relevant fiber bundles. In order to perform group analysis, one last step is necessary: the corresponding bundles of all subjects must be aligned. Therefore, both control and patient images are registered to the ICBM DTI-81 atlas using affine registration. The transformation obtained during this registration is applied to the fiber bundles as seen in the last row of **Figure 1**.

Tract Profiling

Tract profiles are cross sections of the fiber tracts that connect the ROIs specified by the connection matrix generated in preprocessing. For the connections in each ROI pair, a fiber bundle is formed based on the intersections of cross-sectional areas of all subjects' DWI. Then, the medial line of the fiber bundle is computed. Finally, a cross-sectional profile is generated along the medial line so that the distribution of PDDs along each cross section is aggregated separately for each subject group.

[1]TUBITAK 1001, no: 109E081, Ethical board approval: Ankara University Medical College.

[2]A black point is called a *border point* if it is six-adjacent to at least one white point. A black point is called an *end point* if it has exactly one black 26-neighbor. Black point p is simple in $(Z^3,\ 26,\ 6,\ B)$ if and only if all the following conditions hold (Palágyi et al., 2001):

(1) The set $N_{26}(p) \cap (B \setminus \{p\})$ is not empty (p is not an isolated point).
(2) The set $N_{26}(p) \cap (B \setminus \{p\})$. is 26-ected
(3) The set $(Z^3 \setminus B) \cap N_6(p)$ is not empty (p is a border point).
(4) The set $(Z^3 \setminus B) \cap N_6(p)$ is six-connected in the set $(Z^3 \setminus B) \cap N_{18}(p)$.

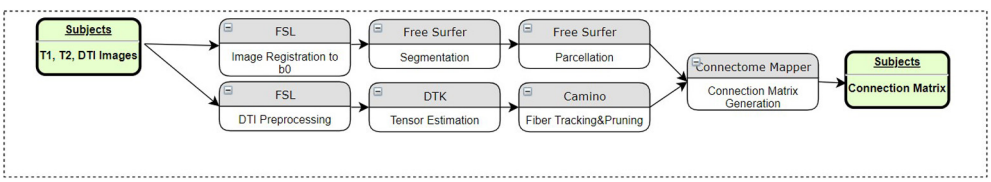


FIGURE 1 | Components of the pre-processing pipeline before TPDS is performed.

Overlapping Fiber Calculation

Overlapping fibers/voxels are calculated across all of the subjects. This is done for each fiber bundle by calculating its maximum overlap. During this process, some specific bundles might be left out as outliers. In **Figure 2**, the overlapping fiber bundle is shown between the two ROIs: thalamus (green) and rostral anterior cingulate (purple). The bundles shown with the yellow, cyan, green, and red colors are marked as outliers and left out of the overlapping area.

The voxel image can be represented as image $\mathcal{P}$ where $\mathcal{P} = (Z^3, m, n, B)$ (Kong and Rosenfeld, 1989). Each element in Z^3 is called a point of $\mathcal{P}$ and each point in $B \subseteq Z^3$ is called a black point and assigned 1. Each point in $Z^3 \backslash B$ is called a white point and assigned 0. m holds black points and n holds white points.

In order to be used in multisubject analysis, adaptation of this definition can be made as follows. For given ROI pairs (i,j), let $\mathcal{P}_0$, $\mathcal{P}_1, \ldots, \mathcal{P}_K$ be a set where K is the number of subjects, and $\mathcal{P}_k$ is the fiber bundle image from subject k. A point in P is assigned as black point if and only if it is also black point for all sets in $\mathcal{P}_0$, $P_1, \ldots, \mathcal{P}_K$ for a given ROI(i,j).

Medial Line Generation

The skeleton of the overlapping bundles is calculated. The curve skeleton is a one-dimensional set which runs through the center of the overlapping bundles in such a way that it preserves the topological properties of the overlapping area. Connectivity conditions are defined as follows. The sequence of points $(x_0, x_1, \ldots, x_n)$ is a j-path of length $n \geq 0$ from the point x_0 to point x_n in a nonempty set of points X if each point of the sequence is in X and x_i is j-adjacent to x_{i-1} for each $1 \leq i \leq n$. The adjacency can be defined as $N_j(p)$ the set of points j-adjacent, to the point p, where $j = 6, 18, 26$. Connectivity can be defined as j-connected if there is a j-path between them in X.

In order to construct the aforementioned skeleton, first of all, curve thinning (Blum, 1967; Kong and Rosenfeld, 1989) is used on P. The medial line of the fiber bundles was generated as depicted in Palágyi et al. (2001). As such, in each iteration, border points of P were deleted until no more deletion was possible. The algorithm is implemented as sequential iterations where each step checks for six subroutines for each of the six-directions that are immediate neighbors of a black point in P. In each iteration, border points are deleted upon satisfying a condition called simple point condition[2]. In this way, the object is shrunk uniformly in each direction. The operation is continued until no more shrinking is possible for each direction. By adding connectivity conditions, the skeleton ends up with the medial line in the near center of the object. In **Figure 2**, the example medial line for the fiber bundle is shown with dark blue.

Finally, the resulting medial line is smoothed by generating a b-spline representation as follows. In order to generate b-spline representation of the medial line, the voxel coordinates on the medial line are represented as data points $\{P_k\}$, $k \in$ MedialLine . A b-spline curve that fits the data is parameterized by $t \in [0, 1]$, where $X(t) = \sum_{i=0}^{n} U_{i,d}(t) Q_i$, the control points Q_i are unknown quantities that have been evaluated using the least-squares fitting method described below:

For n control points $\hat{Q} = \begin{bmatrix} Q_0 \\ Q_1 \\ \vdots \\ Q_n \end{bmatrix}$, and m sample points $\hat{P} = \begin{bmatrix} P_0 \\ P_1 \\ \vdots \\ P_m \end{bmatrix}$, the least-square error function between the b-spline curve and the sample points is the scalar valued function:

$$E(\widehat{Q}) = \frac{1}{2} \sum_{k=0}^{m} \left| \sum_{j=0}^{n} U_{j,d}(t_k) Q_j - P \right|^2$$

To minimize the error function, E, where it is quadratic in the components of $\hat{Q}$, it is a graph of a paraboloid, so it has global minimum that can be found when all its first-order derivatives are 0. The first-order partial derivatives can be written as control points, Q_i

$$\frac{\partial E}{\partial Q_i} = \sum_{k=0}^{m} \left(\sum_{j=0}^{n} U_{j,d}(t_k) Q_j - P_k \right) U_{j,d}(t_k)$$

$$\frac{\partial E}{\partial Q_i} = \sum_{k=0}^{m} \sum_{j=0}^{n} U_{i,d}(t_k) U_{j,d}(t_k) Q_j - \sum_{k=0}^{m} U_{i,d}(t_k) P_k$$

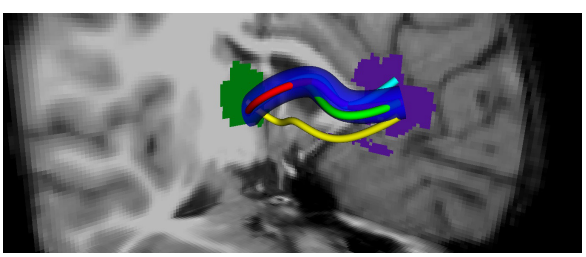


FIGURE 2 | Tract profiling: Generation of an overlapping tract bundle between two ROIs (shown by green and blue) for all of the subjects regardless of the groups.

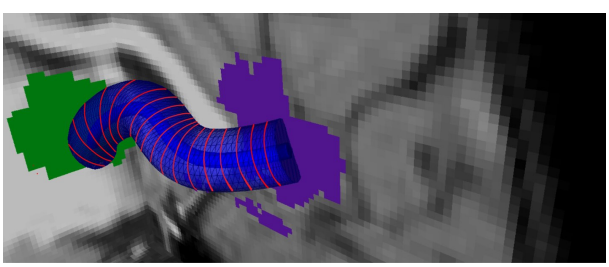


FIGURE 3 | Tract profiling: Representation of the medial line of the overlapping bundle with b-splines and generation tract profiles.

It can be written as $\sum_{k=0}^{m}\sum_{j=0}^{n} a_{k,i}a_{k,j}Q_j - \sum_{k=0}^{m} a_{k,i}P_k$, where $a_{k,i} = U_{k,d}(t_k)$ for $0 \le i \le n$, by setting the partial derivatives to zero vector, and it leads to the system of equations:

$$0 = \sum_{k=0}^{m}\sum_{j=0}^{n} a_{k,i}a_{k,j}Q_j - \sum_{k=0}^{m} a_{k,i}P_k = A^T A\hat{Q} - A^T\hat{P}$$

Where $A = [a_{rc}]$ is a matrix with $m + 1$ rows and $n + 1$ columns.

$$\hat{Q} = (A^T A)^{-1}A^T\hat{P} = \left[(A^T A)^{-1}A^T\right]\hat{P} = X\hat{P}$$

Since A is tridiagonal where it has a contiguous set of upper bands and lower bands, the equation can be solved with the Cholesky decomposition and the vector of control points $\hat{Q}$ can be found.

Since derivative of spline is 1 less order of yet another b-spline where new control points are defined as $Q_i = \frac{p}{u_{i+1+1} - u_{i+1}}(P_{i+1} - P_i)$ from the surface tangent, a normal vector has been computed and cross-sectional areas have been extracted.

Calculation of Tract Cross Sections

The skeleton is sliced with 2-mm regular intervals so that cross-sectional areas that are perpendicular to the b-spline are obtained using normal vectors computed from the surface tangents in **Figure 3**. For each voxel in P that intersects with these cross-sectional areas, PDDs that represent individual subjects are added as tract profiles representing that slice. Hence, for a tract with J slices, there are J tract profiles that contain PDDs which are representative of the subject group. An example tract profile (i.e., a slice with PDDs) from a single subject is shown in **Figure 4**. The PDDs from the subjects for a specific group are aggregated as follows. At each slice, there are fixed number of voxels, and at each voxel, there can be multiple PDDs, each coming from a different subject, depending on whether the subject's tract goes through that voxel or not.

Directional Statistics

Statistical analysis is executed exclusively on areas that are defined by tract profiles eliminates voxel-wise comparison. Hence, misalignment problems no longer exist. Hypothesis testing is conducted only at cross-sectional tract profiles that are separated by 2-mm regular intervals. For the set of PDDs embodied in each tract profile j, a parametric directional statistic distribution is fitted. Through such parametrization, the PDDs of all subjects that fit into the tract profile j are projected onto a sphere.

Watson distribution in **Figure 5** is bimodal and symmetrical around mean direction. Watson distribution assumes that diametrically opposite points have the same probability. Also, the probability density function of axial distributions process antipodal symmetry [i.e., $f(-l,-m,-n) = g(l,m,n)$]. The probability distribution of random vectors that belong to the Watson's family is spherical on a sphere. Directional statistics have been used in the analysis of DTI previously (Schwartzman et al., 2005; Hutchinson et al., 2012), and it has been shown that DTI principal direction analysis using directional statistics can better identify the differences in anatomic structure between populations compared with statistical tests of scalar values such as FA. Both of these studies used Watson distribution to analyze principal directions. On the other hand, Bingham distribution (**Figure 6**) is bimodal and elliptical (Fisher et al., 1993; Cheng et al., 2014). Bingham distribution is free from symmetrical constrains; hence, it provides more advanced distribution fitting options in comparison with Watson distribution.

Watson distribution is defined as follows (Mardia and Jupp, 1999):

$$\text{Watson Distribution } W_p(x;\mu,\kappa) = c_p(\kappa)\, e^{\kappa\left(\mu^T x\right)^2};$$

$$c_p(\kappa) = \frac{\Gamma\left(\frac{p}{2}\right)}{2\pi^{p/2}M\left(\frac{1}{2},\frac{p}{2},\kappa\right)}$$

where x is the unit random vector, μ is the mean vector, is the concentration value, M is Kummer's confluent hypergeometric function, Γ is a gamma function, and p is the dimension of the distribution. To estimate maximum likelihood of this function, we take logarithm. Hence, the log-likelihood function is

$$l(\mu,\kappa \pm x_1,\ldots,\pm x_n) = \kappa\sum_{i=1}^{n}\left(x_i^T\mu\right)^2 - n\log M\left(\frac{1}{2},\frac{p}{2},\kappa\right)$$

$$= n\left\{\kappa\mu^T\bar{T}\mu - \log M\left(\frac{1}{2},\frac{p}{2},\kappa\right)\right\}$$

where $\bar{T}$ is the scatter matrix of the given data. Differentiation with respect to κ gives

$$D_p(\kappa) = \hat{\mu}^T\bar{T}\hat{\mu};\ \text{for } p = 3; = \frac{M(1.5,3.5,\kappa)}{3{*}M(0.5,1.5,\kappa)}.$$

And to find its maximum likelihood estimate, we need a derivative of $D_p(\kappa)$ for $p = 3$

$$D_3^{'} = \frac{M(2.5,3.5,\kappa)}{5M(0.5,1.5,\kappa)} - \frac{1}{9} * \left(\frac{M(1.5,2.5,\kappa)}{M(0.5,1.5,\kappa)}\right)^2.$$

The Newton–Raphson method can be used tfind maximum values for $D_p(\kappa)$ and the biggest eigenvalue of scatter matrix, t_1, for a bipolar distribution or t_3 for a girdle distribution.

Bingham distribution is defined as a trivariatnormal distribution on a unit sphere. Different from Watson distribution, it has three orthogonal directions as μ_1, μ_2, μ_3 and concentration values (κ_n) for each orientation vector (Watson and Williams, 1956).

Concentration values define the dispersion of the distribution, where

(1) $\kappa_1 = \kappa_2 = 0$ results in a spherical distribution of axes.
(2) $\kappa_1 = \kappa_2 \ll 0$ results in a symmetric bipolar distribution.
(3) $\kappa_1 < \kappa_2 \ll 0$ results in an asymmetric bipolar distribution.
(4) $\kappa_1 \ll \kappa_2 < 0$ results in an asymmetric girdle distribution.
(5) If $\kappa_1 \ll 0$ and $\kappa_2 = 0$, then Watson distribution is obtained.

The probability distribution function of Bingham distribution is defined as follows (Bingham, 1974):

$$\text{Bingham Distribution } B_p(x;K) = c_p(K)\, e^{x^T K x};$$

$$c_p(K) = \frac{\Gamma\left(\frac{p}{2}\right)}{2\pi^{p/2} F\left(\frac{1}{2}, \frac{p}{2}, K\right)}$$

where x is the unit random vector, K is the 3×3 orthogonal orientation matrix with concentration values, F denotes the confluent hypergeometric function of matrix argument, Γ is the gamma function, and p is the dimension of the distribution. For a given random sample $\pm x_1, \ldots, \pm x_n$, the log-likelihood function can be written as:

$$l(K; \pm x_1, \ldots, \pm x_n) = n \left\{ \log tr\left(A\bar{T}\right) - \log F\left(\frac{1}{2}, \frac{p}{2}, K\right) \right\}$$

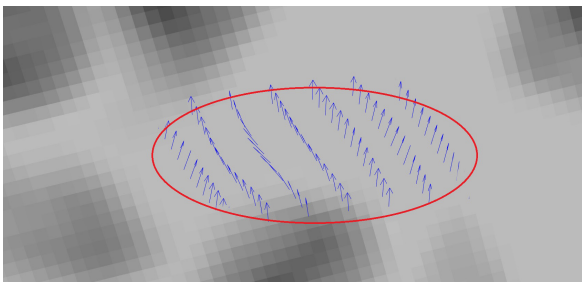


FIGURE 4 | Tract profiling: Illustration of the PDDs from a single subject in a sample tract profile.

We can write K and $\bar{T}$ in polar form as $= UKU^T$, $\bar{T} = VtV^T$ with U and V being orthogonal. $K = \text{diag}(\kappa_1, \ldots, \kappa_p)$ and $t = (\bar{t}_1, \ldots, \bar{t}_p)$, where $\kappa_1 \geq \ldots \geq \kappa_p$ and $\bar{t}_1 \geq \ldots \geq \bar{t}_p$. As suggested by Bingham himself, the following approximations can be used.

For the bipolar case:

$$d = \bar{t}_2 - \bar{t}_3, \quad s = \bar{t}_1 + \bar{t}_2, \quad \kappa_0 = -D_3^{-1}(\bar{t}_1)$$

$$\kappa_1 \approx 0, \kappa_2 \approx \kappa_0 + \delta, \quad \kappa_3 = \kappa_0 - \delta$$

For the girdle case:

$$d = \bar{t}_1 - \bar{t}_2, \quad s = \bar{t}_1 + \bar{t}_2, \quad \kappa_0 = -D_3^{-1}(\bar{t}_3)$$

$$\kappa_1 \approx 0, \ \kappa_2 = -2\delta, \ \kappa_3 = \kappa_0 - \delta$$

where $\delta = \frac{2d\kappa_0}{s(\kappa_0 - 1.5) + 1}$

After parametric representation through either Watson or Bingham distribution, two group of subjects can be compared by using an eclipse of confidence defined by the p value. For fitting a single group's data, the mean direction vector of the group is computed. If it lies inside the eclipse of confidence of the targeted distribution, then the null hypothesis is likely, justifying a reasonable fit to the associated directional distribution. On the other hand, if the confidence ellipse around the mean direction does not overlap for a given confidence level, then the null hypothesis is unlikely, rejecting the fit. For two groups, the case with different means is indicated by separated cones of confidence, which in turn indicates significant differences. On the other hand, overlapping cones of confidence indicate insignificant differences, hence acceptance of the null hypothesis.

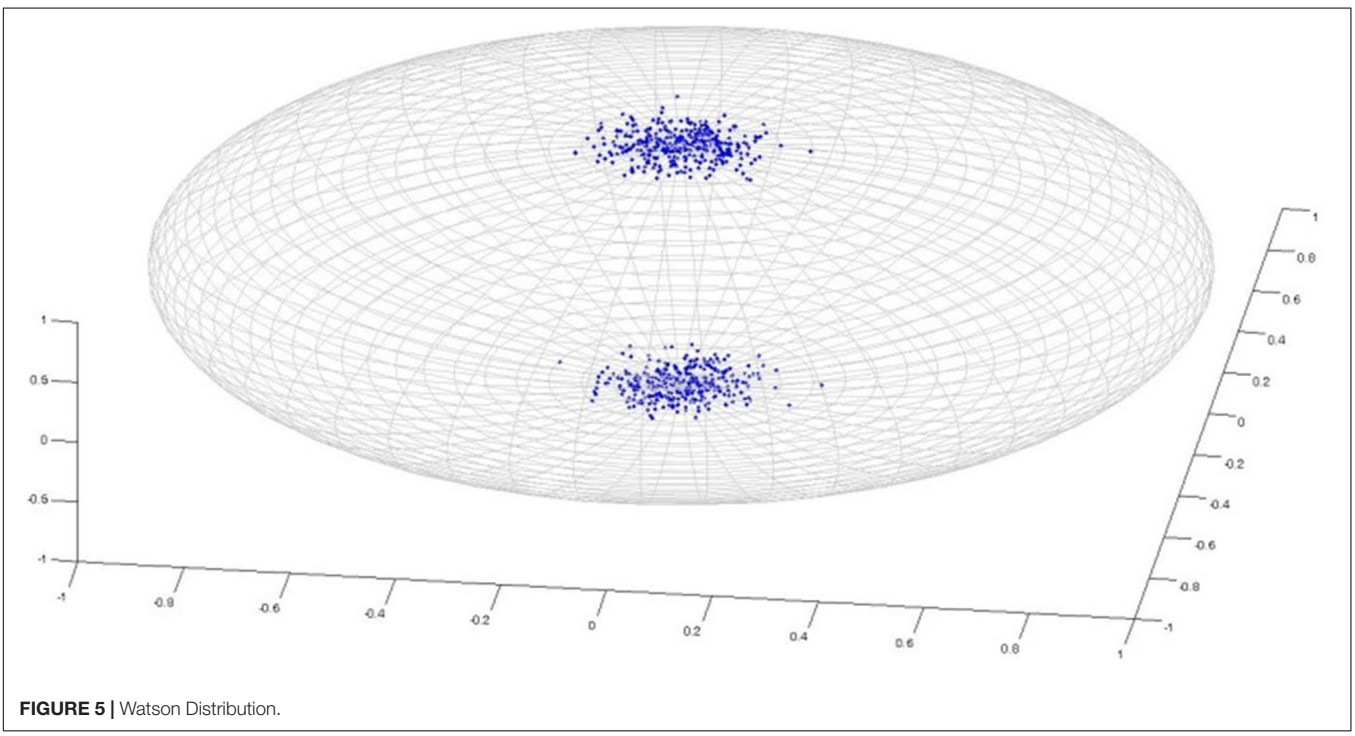


FIGURE 5 | Watson Distribution.

An example distribution is provided in **Figure 7**, for two groups of subjects, for representation with the Bingham distribution.

Details of the eclipse of confidence can be given as follows. The maximum likelihood estimates of concentration parameters κ_1, κ_2 can be obtained from maximizing the log-likelihood function, where w_n are the eigenvalues of the principal eigenvector of the orientation matrix:

$$F = -N\log(4\pi) - N\log d(\kappa_1, \kappa_2) + \kappa_1 w_1 \kappa_2 w_2$$

Maximum likelihood estimators of κ_1, κ_2 in the Bingham distribution for given eigenvalues w_1,w_2 can be estimated as calculated by Mardia and Zemroch (1977).

The confidence ellipse around the mean direction within the specified percentage (%) of the estimated concentration values of distribution as

$$e_{\%}^{mn} = \sqrt{\left[\frac{X_{\%}^2}{2N(\Delta_{mn})}\right]} \text{for} \Delta_{mn} = (\kappa_m - \kappa_n)(w_m - w_n) \text{and} X_{\%}^2$$

is the chi-squared value for two degrees of freedom and % is p value for confidence interval.

For $p = 0.01$ and having $\kappa_3 = 0$ (Fisher et al., 1993) ends up with the semi-axes of the confidence eclipse about the mean direction associated with w_3 as below:

$$e_{32} = -1.517\frac{1}{k_2 N(w_3 - w_{2)}} \text{and} e_{31} = -1.517\frac{1}{k_1 N(w_3 - w_{1)}}$$

Performance Analysis

We have conducted two performance tests to analyze the effectiveness of the proposed method. First, we have analyzed the models generated by TPDS in comparison with VBA and TBSS. Hereby, we have adapted directional statistics to TPDS, TBSS, and VBA to compare their overall efficiency in representing vector-based statistical models. This test aimed to show the efficiency of tract profiling over voxel-based and skeleton-based analysis. Second, we have applied the full TPDS algorithm to the two subject populations (i.e., MDD versus healthy controls) and compared the results with NBS. This test aimed to show the efficiency of combining tract profiles with directional statistics over conventional methods. In this test, the effects of fiber length in estimating group differences were also evaluated.

Analysis of the Strengths of VBA, TBSS, and TPDS in Tract Modeling

In this test, we used a single group (i.e., healthy subjects). The statistics were derived using three different methods, VBA, TBSS, and TPDS, only on white matter areas—not using GM ROIs. As seen in **Figure 8**, the white matter areas that have been segmented using FreeSurfer are mapped to ICBM DTI-81 atlas (Mori et al., 2008) to allow for intersubject data aggregation. For VBA analysis, the atlas-based white matter areas are overlayed for all subjects for further processing. For TPDS analysis, tract profiles are generated from the atlas mappings of all subjects. In TBSS, before performing atlas mapping, skelotonized areas are generated from individual subject tracts. The rest of the data processing pipeline is the same for all three methods. At the first step, for each WM ROI, based on which method is used for defining the tract, PDDs are generated. Then these PDDs are parametrically modeled by two separate directional distributions, namely Bingham and Watson. Finally, in the last step, several PDDs are generated to represent the entire group using the newly developed parametrical models, and goodness of fit is computed to evaluate how good the chosen model is.

PDD Generation

For each subject, primary diffusion directions are extracted for each voxel inside the given WM area using the primary eigenvector of the diffusion tensor. The WM area differs based on the chosen representation. In VBA, the WM area is extracted based on segmentation of the specific WM ROI. In TBSS, it is based on the skeleton of the tract in the WM ROI. In TPDS, it is embodied within each tract profile that composes the entire tract in the WM ROI. Aggregated data from all subjects compose the data to be fitted for each WM area.

Distribution Fitting

Watson and Bingham distributions were fitted to model each tract using the maximum likelihood method. For each tract, the parameters of the theoretical model were estimated from the pdf at hand. Then this theoretical probability density function was evaluated iteratively using synthetic random vector data for a total of 700 vectors that were almost uniformly distributed along a sphere. Finally, the difference between the estimated pdf and the random pdf is tested for null hypothesis.

Goodness of Fit Testing

Pearson's chi-square tests have been used for goodness of fit tests to evaluate whether the observed frequency distribution differs from the theoretical distribution. Comparison of distributions is done using ANOVA and the chi-square test statistics was also used for each ROI.

In order to apply Pearson's chi-square tests to check whether the observed frequency distribution differs from a theoretical distribution, the following steps are applied on the original data and synthetic random vector data and the respective models.

(1) For Watson distribution, the sample mean direction, $\bar{R}$, has been evaluated as a regular vector sum of the vectors under a population of vectors. The mean direction is a unit vector that is in the same direction with R: $\bar{x} = \frac{\sum_i x_i}{R}$, $\bar{y} = \frac{\sum_i y_i}{R}$, $\bar{z} = \frac{\sum_i z_i}{R} R$.
(2) For Bingham distribution, the axis of moment of inertia of sample, $\bar{t}$, has been evaluated using the scatter matrix of distribution S. For the bipolar case, it is the biggest eigenvector, and for girdle case, it is the smallest eigenvector.
(3) The transformations $\bar{\theta}$, $\bar{\phi}$ have been evaluated in order to shift either $\bar{R}$ or $\bar{t}$. to positive z-axis.
(4) The transformation has been applied to original and synthetic data.

(5) The angle θ has been calculated as the angle between the positive x-axis and the projected vector on the x–y plane: $0 < \theta < 2\pi$.
(6) The observed frequencies and the excted frequencies were $\theta_1 < \theta < \theta_2$, where the number frequency bins is 50 and $\theta_2 - \theta_1 \approx 7.2^o$.

Analysis of the Group Difference Maps Generated by NBS and TPDS

In this part, the proposed framework will be applied to test for differences of fiber tract profiles between MDD patients and control subjects. Based on the same fiber tracts and connectivity matrix for healthy volunteers, comparisons will be made with the results of the network-based statistics. For this purpose, we used the 83 × 83 connectivity matrix generated at the end of data preprocessing by the Connectome Mapper (**Figure 1**).

In NBS, for each group, each pairwise association (i,j) between ROI i and ROI j is treated separately. First, Fisher's r-to-z transform has been applied to ensure normality. Then, the test statistic of interest—which is the normalized number of fiber bundles—is compared between the groups using t-statistic. In order to correct for multiple comparisons, permutation testing was used to select the p value controlled for the FWE for each connected component. For each permutation, the same threshold is applied to define a set of suprathreshold links of connected components. Suprathreshold and the number of permutations were set according to the default parameter settings of NBS with corrected $p < 0.005$.

In TPDS, the following procedure is repeated for each possible connection between distinct ROI pairs (i.e., 83 × 83 times divided by 2). Tract profiles between each ROI i and ROI j are extracted for the healthy and MDD groups. Then for each slice in the tract profiles, significance is tested with a threshold value of $p < 0.005$. If there are n contiguous slices that satisfy this, it is indicated that the connection between ROIs i and j is significantly different between the control and patient groups. It is possible that there are multiple clusters of n contiguous slices that satisfy this condition. In order to reflect this information, we prepared a new 83 × 83 connectivity matrix, which contained the number of significantly different clusters between the two groups that are compared. Therefore, the difference map that is achieved through TPDS reflects a weighted graph, weight being the number of significantly different clusters between the two groups for that particular i to j connection. The more the number of significantly different n contiguous slices, the more the weight of the difference map.

Selection of n must be done according to a criterion related to the plausible tract lengths. In order to eliminate premature tract termination that result from low SNR and low pathway anisotropy (Behrman-Lay et al., 2015), 10 mm is the shortest tract length to be considered. Since DTI image has 2.2 mm spacing, choosing n as 4 satisfies this constraint. In other words, at least four consecutive cross-sectional areas must be found within a fiber bundle where the PDD of each cross-sectional area belongs to significantly different Bingham distributions for the control and MDD groups.

RESULTS

The results of the performance tests that we performed to investigate the effectiveness of TPDS are as follows.

Comparison of VBA and TBSS With TPDS Using Directional Statistics

As can be seen in **Table 1**, among VBA, TBSS, and TPDS, the best fitted distribution is more representative in TPDS because the goodness of fit scores are better according to p values. In addition, based on the results of TPDS, the Bingham distribution is reported to be more favorable than the Watson distribution because only 2 out of 48 white matter tracts are represented better with Watson. Obviously, it is evident that TPDS is a better alternative to represent tracts in comparison with VBA and TBSS, because it favors a more parametrical fit to the entire set of fiber tracts.

A close inspection of **Table 1** reveals that in terms of representing a given WM tract parametrically, TBSS is superior to VBA, and TPDS is superior to TBSS. It is evident that VBA contains more noise than TBSS and TPDS, because it contains the entire WM area from all subjects. Due to high noise, VBA fails to represent some of the tracts parametrically. On the other hand, TBSS is better than VBA, because it removes the areas—hence the noise associated in these parts—that lie outside the fiber bundles which constitute the skeleton. However, TBSS is not better than TPDS, because it smooths out the tracts while forming the skeleton and loses specificity. Overall, the tract profiles computed in TPDS are selective in choosing representative samples of the DWIs that are more informative, because outliers are removed while computing the medial line. Since the data points all belong to the same tract and on the same cross section over the medial line, very similar diffusion properties are expected for each analysis point. This tends to eliminate all negative effects of misalignment of images and partial volume effect. Due to this property, the computational effectiveness of TPDS is higher than other methods, because the model can be decided with much less number of data points.

The advantage of the Bingham distribution might be explained through the ease of fitting a girdle distribution in comparison with fitting a homogeneous mean direction distribution. The girdle distribution allows for more parameters; hence, it makes the development of a more general model possible. Furthermore, the computational accuracy of the Bingham distribution is better because the tracts represented with this distribution fit to the PDD of the actual tracts with a smaller p value.

Comparison of the Group Differences in Connectivity Maps Using Network-Based Statistics and TPDS

In NBS, with corrected $p < 0.005$, seven regions and eight connections have been observed to contain lower FA in MDD. Particularly, the connections in the right hemisphere and between the superior frontal cortex and rostral/caudal components of the anterior cingulate cortex, caudate, and inferior

TABLE 1 | Comparison of voxel-based analysis (VBA) and tract-based spatial statistics (TBSS) with tract profiling and directional statistics (TPDS) (VBA and TBSS have been adapted to run directional statistics).

WM tract	VBA model (*p* value)	TBSS model (*p* value)	TPDS model (*p* value)
Middle cerebellar peduncle	No fit (0.803)	No fit (0.425)	Bingham (0.021)
Pontine crossing tract	No fit (0.092)	Bingham (0.043)	Bingham (0.004)
Genu of corpus callosum	Bingham (0.030)	Bingham (0.032)	Bingham (0.007)
Body of corpus callosum	Bingham (0.001)	Bingham (0.001)	Bingham (0.031)
Splenium of corpus callosum	Bingham (0.046)	Bingham (0.036)	Bingham (0.045)
Fornix (column and body of fornix)	No fit (0.707)	No fit (0.135)	Bingham (0.017)
Corticospinal tract R	No fit (0.067)	No fit (0.087)	Bingham (0.048)
Corticospinal tract L	No fit (0.541)	Bingham (0.041)	Bingham (0.025)
Medial lemniscus R	No fit (0.706)	Watson (0.046)	Bingham (0.036)
Medial lemniscus L	No fit (0.278)	No fit (0.078)	No fit (0.090)
Inferior cerebellar peduncle R	Watson (0.019)	Watson (0.037)	Bingham (0.016)
Inferior cerebellar peduncle L	No fit (0.970)	No fit (0.570)	Bingham (0.019)
Superior cerebellar peduncle R	Watson (0.032)	Bingham (0.042)	Bingham (0.045)
Superior cerebellar peduncle L	Watson (0.026)	Bingham (0.044)	Bingham (0.012)
Cerebral peduncle R	No fit (0.064)	Bingham (0.044)	Bingham (0.023)
Cerebral peduncle L	No fit (0.078)	Bingham (0.032)	Bingham (0.022)
Anterior limb of internal capsule R	Watson (0.030)	No fit (0.079)	Watson (0.023)
Anterior limb of internal capsule L	Bingham (0.002)	Bingham (0.038)	Watson (0.025)
Posterior limb of internal capsule R	Watson (0.014)	No fit (0.067)	Bingham (0.008)
Posterior limb of internal capsule L	Bingham (0.017)	Bingham (0.033)	Bingham (0.033)
Retrolenticular part of internal capsule R	Watson (0.034)	No fit (0.074)	No fit (0.083)
Retrolenticular part of internal capsule L	Watson (0.015)	No fit (0.065)	No fit (0.106)
Anterior corona radiata R	No fit (0.278)	Bingham (0.012)	Bingham (0.045)
Anterior corona radiata L	Bingham (0.0012)	Bingham (0.002)	Bingham (0.001)
Superior corona radiata R	Watson (0.043)	Bingham (0.009)	Bingham (0.001)
Superior corona radiata L	No fit (0.165)	Bingham (0.035)	Bingham (0.019)
Posterior corona radiata R	Watson (0.002)	Bingham (0.017)	Bingham (0.002)
Posterior corona radiata L	Watson (0.001)	Bingham (0.019)	Bingham (0.006)
Posterior thalamic radiation R	Bingham (0.003)	Bingham (0.002)	Bingham (0.024)
Posterior thalamic radiation L	Bingham (0.006)	Bingham (0.002)	Bingham (0.009)
Sagittal stratum R	No fit (0.188)	No fit (0.488)	Bingham (0.032)
Sagittal stratum L	No fit (0.065)	No fit (0.265)	Bingham (0.047)
External capsule R	Bingham (0.006)	Bingham (0.006)	Bingham (0.001)
External capsule L	Bingham (0.001)	Bingham (0.001)	Bingham (0.001)
Cingulum (cingulate gyrus) R	Bingham (0.015)	Bingham (0.033)	Bingham (0.003)
Cingulum (cingulate gyrus) L	Bingham (0.002)	Bingham (0.001)	Bingham (0.001)
Cingulum (hippocampus) R	Watson (0.004)	Bingham (0.004)	Bingham (0.001)
Cingulum (hippocampus) L	No fit (0.118)	Bingham (0.019)	Bingham (0.041)
Fornix (cres)/stria terminalis	Bingham (0.04)	No fit (0.050)	Bingham (0.009)
Fornix (cres)/stria terminalis	Bingham (0.012)	No fit (0.128)	Bingham (0.005)
Superior longitudinal fasciculus R	Watson (0.002)	Bingham (0.043)	Bingham (0.021)
Superior longitudinal fasciculus L	Watson (0.025)	Bingham (0.040)	Bingham (0.011)
Inferior fronto-occipital fasciculus R	Bingham (0.025)	Bingham (0.008)	Bingham (0.003)
Inferior fronto-occipital fasciculus L	Bingham (0.004)	Bingham (0.062)	Bingham (0.002)
Superior fronto-occipital fasciculus R	Bingham (0.003)	Bingham (0.018)	Bingham (0.001)
Superior fronto-occipital fasciculus L	Bingham (0.044)	Bingham (0.026)	Bingham (0.006)
Uncinate fasciculus R	No fit (0.483)	No fit (0.091)	Bingham (0.003)
Uncinate fasciculus L	No fit (0.896)	Bingham (0.039)	Bingham (0.002)
Tapetum R	No fit (0.595)	No fit (0.092)	Bingham (0.092)
Tapetum L	No fit (0.535)	No fit (0.103)	Bingham (0.004)

parietal cortices had lower FA in MDD. These connections are shown in **Figure 9** as green lines.

In TPDS, significantly different connections between the healthy and MDD groups are seen in **Figure 9** as red lines. The thickness of the lines reflects the weights or in other words the number of cross-sectional areas above the threshold n (e.g., A weight value of 1 indicates that there exists only one slice cluster with significantly different *n* contiguous tract

profiles, whereas a weight value of 6 indicates that there exist 6 disjoint clusters of n contiguous tract profiles that are significantly different). The right hemisphere differences reported by NBS, namely the frontal (superior frontal and rostral middle frontal) and medial (caudal and rostral anterior cingulate), are also detected by our method. But, additionally, TPDS revealed differences between the healthy and MDD populations in limbic, temporal cortex, occipital cortex, and hippocampal connections, as well as a few left hemisphere areas such as the amygdala, hippocampus, and thalamus.

The strength of the tract profile structure lies in the reduction of the misalignment problem. Furthermore, observations of the

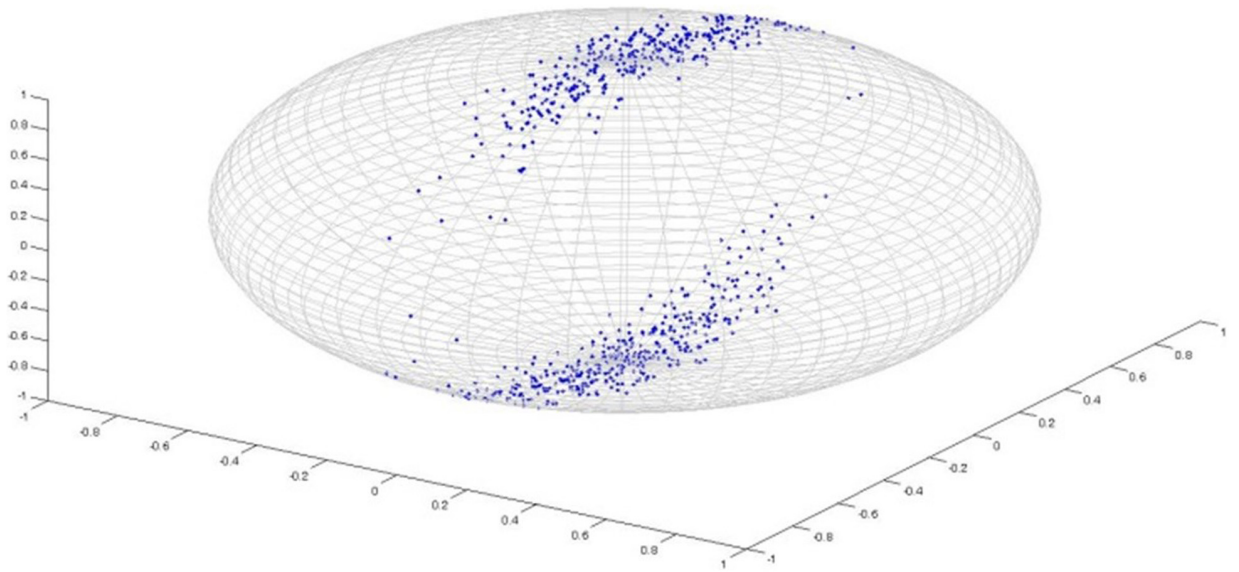


FIGURE 6 | Bingham Distribution.

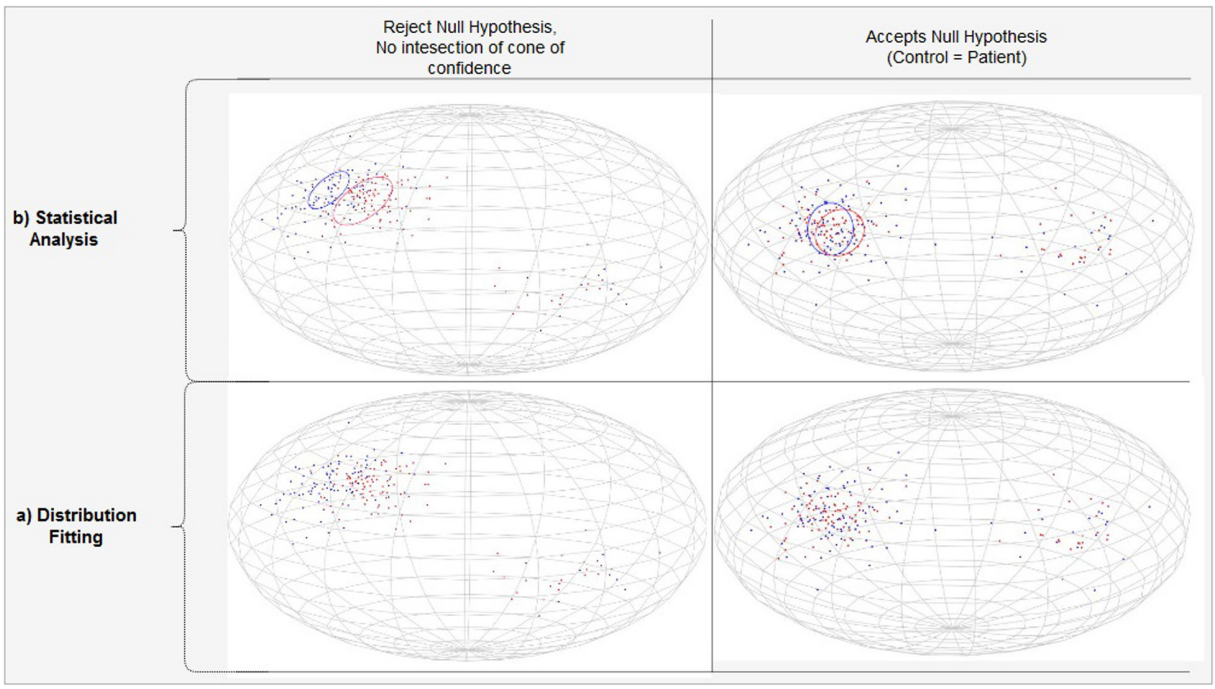


FIGURE 7 | PDD projections modeled by the Bingham Distribution. **(A)** Separated (Left) versus overlapping (Right) vector projections of PDDs on unit sphere for two different subject groups shown with blue and red. **(B)** Statistically significant (Left) versus insignificant (Right) differencs between populations.

directional changes become more specific because contributions of the local changes can be reported along the tract not by the contribution of isolated voxels but by several slices across the two ROIs. Therefore, the proposed directional statistics comparison is expected to be a superior differentiator for especially long tracts.

In order to verify this, the following analysis has been done. Using TPDS, for each tract connecting 83 different regions, the z-score of each length is plotted against the z-score of the number of significantly different profile slices. For this purpose, the maximum overlapping shape (skeleton) is used. When regression lines are fitted to investigate the relationship with tract length and the number of different clusters, it is seen that the likelihood of finding clusters of voxels that differ in long tracts increase with respect to path length. This has been also tested using a linear regression model, where it has been found that the z-score of tract length significantly correlated with the z-score of the number

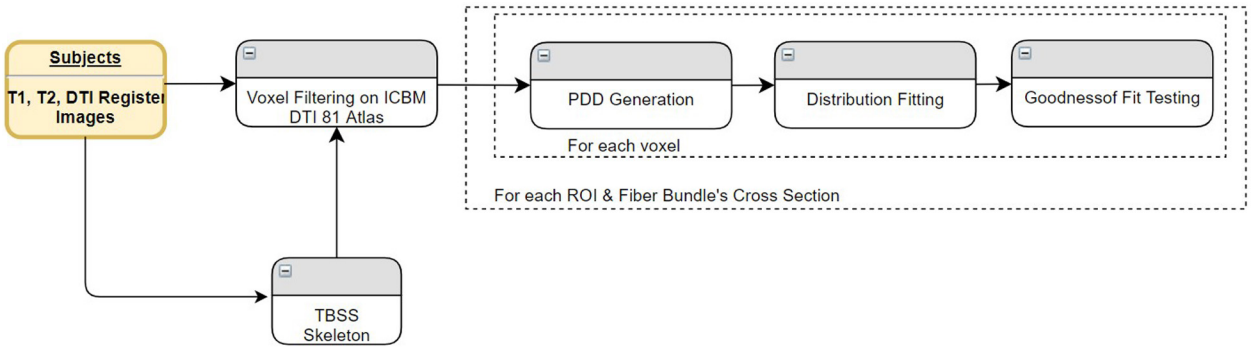


FIGURE 8 | Comparison of VBA/TBSS with TPDS data processing pipeline.

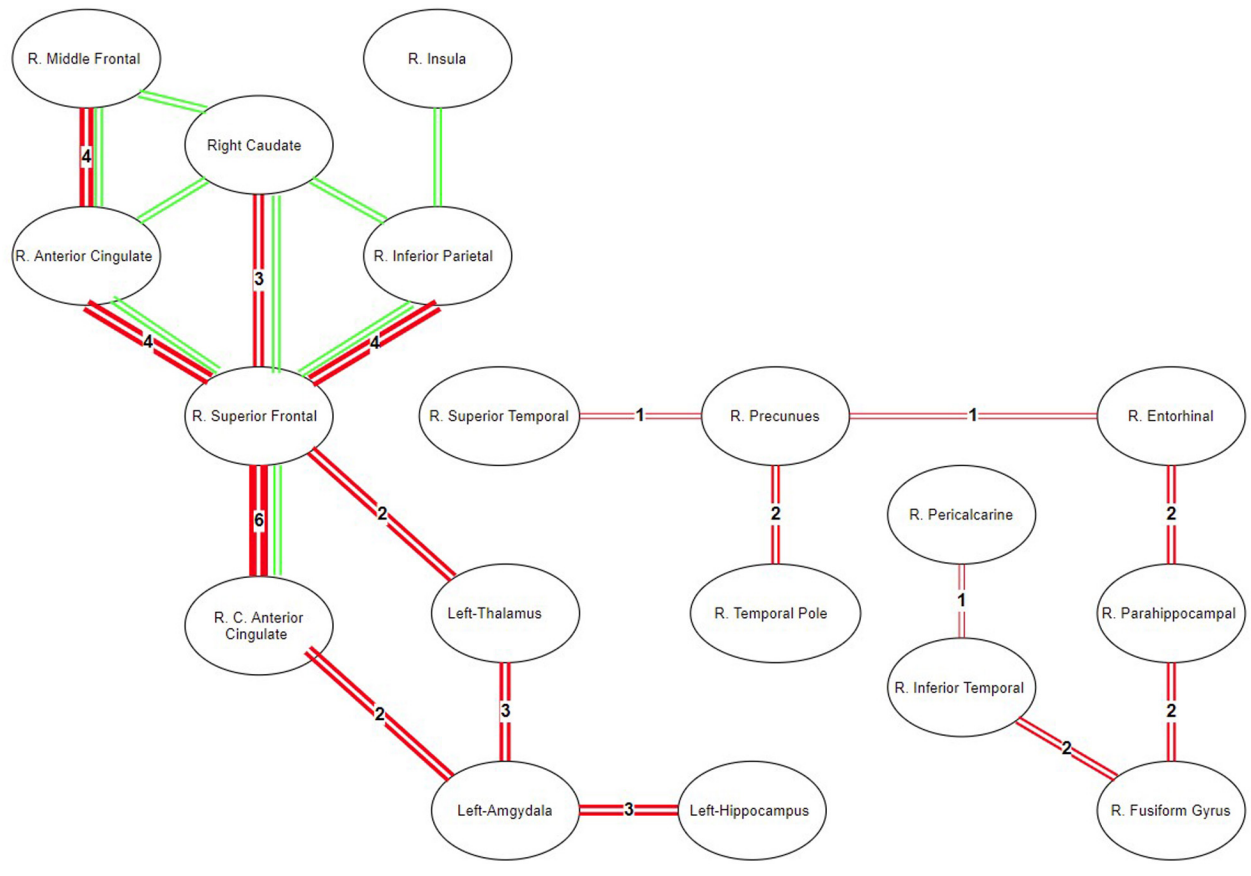


FIGURE 9 | Map of ROIS with statistically different connectivity between control and patient groups. Green lines represent the common connections that are found different between the groups using NBS. Red lines represent the significantly different connections detected by the directional statistics using tract profiling.

of significantly different profile slices ($p < 0.05$, adjusted R^2: 0.00162) as seen in **Figure 10**. Although the effect size is small, we can indicate that TPDS is a powerful method to find differences in two populations, especially as the tract lengths get longer.

DISCUSSION

In this study, we proposed a novel framework for WM fiber connectivity analysis using TPDS. In contrast with other group studies (Goodlett et al., 2009) that are based on FA values, directional statistics deals with compact Riemannian manifolds, which allow observations regarding local diversities of principal diffusion directions of voxels in different groups of subjects.

Comparison of TPDS With Other Techniques Used in Analysis of Groups of DWI

Our pipeline implementation can be regarded as quantitative tractography. We analyze diffusion properties on the exact tracts and derive the statistics over sample points taking neighborhood cells into consideration. A similar method has been offered by Corouge et al. (2006) where diffusion properties along the fiber tracts, called fiber property profiles, are extracted. In that study, fiber tract parameterization was based on arc length parameter, starting from each fiber's intersection with an "origin" plane. Goodlett et al. also proposed a similar tract profiling approach, where diffusion properties are calculated along the tract for each fiber bundle (Goodlett et al., 2009). Our method introduces three main improvements to these quantitative tractography methods. First, we are not just limiting the method with known anatomical fiber bundles but can derive statistics from any pair of connected gray matter areas. Second, we have introduced skeletonization and pruning to allow for applying statistics only within common areas across the groups. Third, we introduced vector analysis using directional statistics over scalar analyses such as FA, MD, etc.

There exist other methods which use directional statistics in DTI (Schwartzman et al., 2005; Hutchinson et al., 2012). However, these methods analyze group differences based on ROIs, not fiber tracts, ignoring the underlying connectivity. We have devised the tract profiling algorithm to operate on relevant voxels among the fibers that connect each ROI obtained from fully automatic brain segmentation and parcellation. Local registration errors are reduced after calculating cross-sectional area of the fibers and finding medial lines (i.e., profiles) to continue tract analysis. Afterwards, Bingham distribution, which is the most general form of directional distribution, is used for tract-based directional analysis, ensuring minimum parametric assumptions about the dataset. To the best of our knowledge, this approach has not been implemented in group analysis of DTI before.

Neurite orientation dispersion and density imaging (NODDI) is a novel neurite imaging and analysis framework and provides sensible neurite density and orientation dispersion estimates. Unlike FA, NODDI analyzes density and orientation dispersion separately. NODDI uses orientation distribution function (ODF), defined as Watson distribution which constrains the dispersion about the dominant orientation (Zhang et al., 2012). However, Bingham distribution fits better to diffusion properties, in comparison with Watson. Bingham-NODDI extends the NODDI

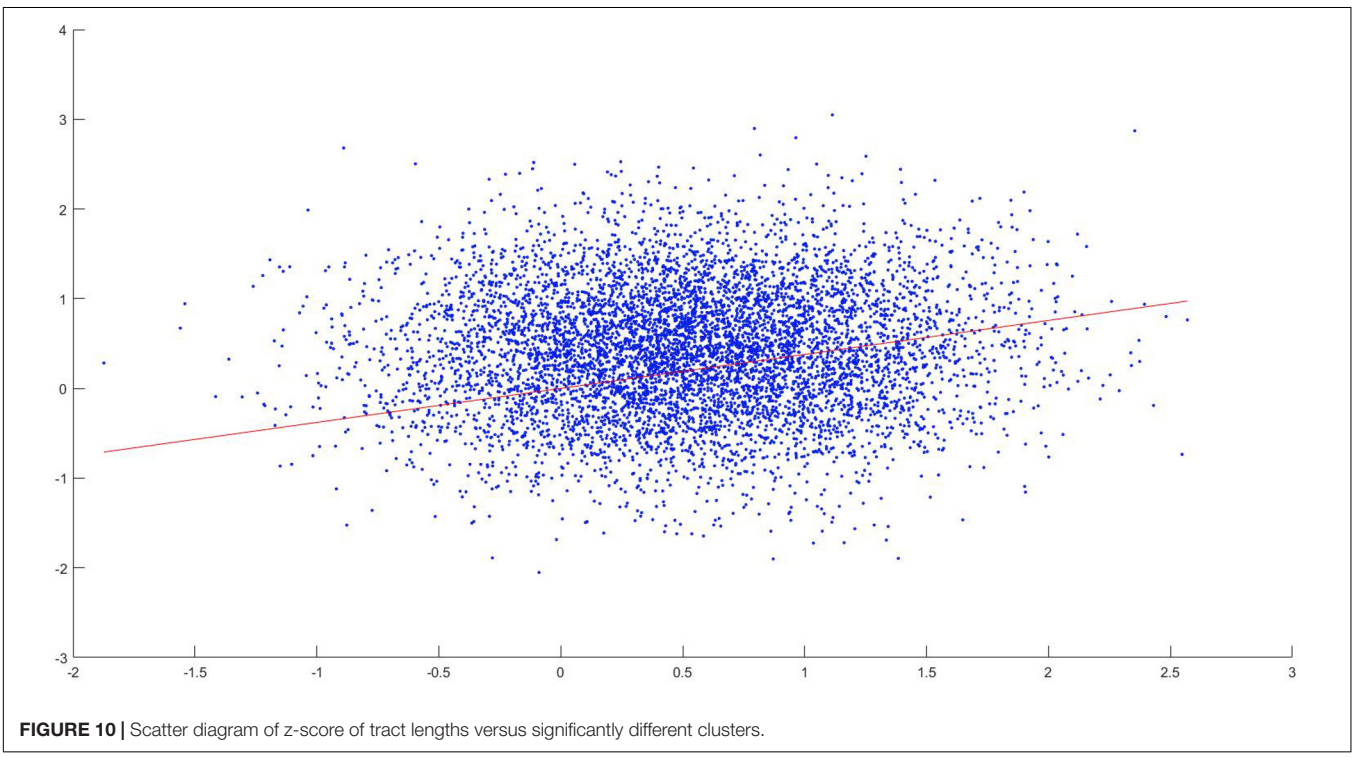


FIGURE 10 | Scatter diagram of z-score of tract lengths versus significantly different clusters.

method by generalizing it with Bingham distribution to cover anisotropic orientation dispersions of neurites (Tariq et al., 2016). Similarly, in our study, we found that modeling ODF using Bingham distribution explains the data better regardless of the tract identification, be it through VBA, TBSS, or our method, TPDS. A major difference between our approach and NODDI is in the estimation of the dispersion. The modeling we used to implement the Bingham distribution estimates dispersion in the vicinity of the dominant orientation, separately for the primary and secondary dispersion orientations. This eliminates the key limitation of NODDI, failing to model complex neurite configurations such as those arising from fanning and bending axons. On another front, just like ours, orientation dispersion (ODI) generated by the NODDI method can be also used with the TBSS method instead of FA metric (Timmers et al., 2016; Taoka et al., 2020). In this aspect, the main difference between our method and NODDI is in extending fiber dispersion along the tracts that connect two ROIs. By allowing such extension, our method enables using fiber dispersions as track characteristics and analyzing disease-related effects on connectivity of the tracks rather than the voxel.

We have demonstrated that in addition to scalar diffusibility changes, analyzing principal diffusion directions along a tract detects local changes better than scalar values. The strength of the directional statistics-based analysis we proposed lies in its applicability to TBSS and VBA as well; it is not limited to tract profiles.

Voxel-based analysis needs to register the subject's images to a common coordinate frame. However, the fiber tracts do not accurately align during this process due to variation in tract size and shape. Especially, long-range fiber tracts contain more shape variation across subjects (Wassermann et al., 2011), so they are more prone to such misalignment. This problem is still valid for TBSS because even the voxel skeletons do not ensure that all relevant voxels correspond to the same tract (de Groot et al., 2013).

In directional statistics, the misalignment problem though the tract becomes more critical compared with scalar statistics like FA. As seen from the results of the first set of performance tests, tract profiles are superior structures for resolving the shape differences in comparison with VBA and TBSS, because tract profiles are better in terms of fitting a model to PDD vectors. We investigated the goodness of fit characteristics of VBA and TBSS, respectively, on all WM areas and on skeletonized WM areas using directional statistics. We found that several tracts in VBA and TBSS are rejected to fit to the most general Bingham distribution which contains minimum assumptions about the data. In comparison when tract profiling is used, most tracts could be fit parametrically, except a few. A parametrical model is advantageous in data processing, since it facilitates population-based comparisons.

The aforementioned tests also show how directional statistics can be adapted to the widely used analysis methods such as TBSS or VBA. Instead of FA values, PDD vectors can be used over each voxel within the skeleton. FA metric uses eigenvalues of the underlying diffusion characteristics of the voxel and defines only the amount of diffusion asymmetry where PDD uses the first eigenvector of the diffusion characteristic. The FA metric is sensitive to the underlying fiber architecture and correlates with PDD changes in disease conditions. However, FA does not have direction property. Different orientations might result in the same FA value simply because orientational changes of the diffusion property of the voxel might not end with FA changes, when there is a difference in eigenvector orientation but not its value. So, the FA metric is not as sensitive as PDD in detecting diffusion characteristic differences along the fiber track. As can be seen in **Table 1**, Bingham distribution fits better to describe the differences in the majority of white matter tracks. Further studies should be conducted to ease adaptation of directional statistics to TBSS skeletons and also to resolve issues related to the multiple comparison problem.

PDD analysis using directional statistics is not a summary statistics of each track but a measurement of diffusional properties of the fiber bundle connecting a pair of ROIs. The statistics of each voxel along the fiber track are summarized by many points using directional statistics along the fiber bundle. Fiber bundle skeletonization and normalization of PDD over tract cross sections allows for error correction and noise cancelation that might arise from tractography artifacts or misalignment. This should also be valid for trajectory changes of tracts under disease-related conditions, as long as a prominent disfiguration or an abnormal morphological change caused by a tumor deviation does not severely divert the alignment of the fiber bundles. In such a case, a lot of false positives may affect the model along the fiber bundles, hindering the correct estimation of PDDs along the actual but diverted tract.

During the second set of performance tests, the results of TPDS and NBS are compared to see whether these methods report the differences between the healthy and MDD populations consistently. We found that most of the right hemisphere-specific connectivity differences reported earlier in MDD have been detected by both of these approaches. The results are much more consistent among the shorter tracts such as frontal connections of the anterior cingulate. However, TPDS reveals additional connectivity differences mainly among longer tracts such as those between temporal and occipital cortex as well as those that contain areas with low FA values and higher crossing fibers such as the amygdala, hippocampus, and thalamus. Another strength of TPDS is due to its revelation about weights, which indicate the amount of difference between the subject populations along the tracts.

These findings are also consistent with MDD models proposed by Drevets et al. (2008) and Mayberg (2003) where MDD can be defined through a limbic–cortical dysregulation model. In this model, the limbic–thalamo–cortical (LTC) circuits, involving the amygdala, thalamus, and orbital and medial PFC, and the limbic–cortical–striatal–pallidal–thalamic (LCSPT) circuits are mainly the affected areas. These connections are found be affected both using NBS and TPDS. Additionally, TPDS revealed

temporal, parietal, and occipital cortex connections that are different in MDD. Mainly the differences on inferior fronto-occipital tracts can be also supported by other DTI studies that report significantly decreased FA values among MDD patients (Cheng et al., 2014).

The ROIs reported to have statistically significant connectivity differences in MDD versus healthy participants are consistent with the two well-known lateralization models of emotion. According to the right hemisphere hypothesis, the right hemisphere is dominant in processing emotions (Alves et al., 2008). On the other hand, the valence hypothesis posits that the left hemisphere processes positive (or approach-related) information, but the right hemisphere processes negative (or avoidance-related) information (Alves et al., 2008). Within the context of MDD, hypoactivity in the left hemisphere fronto-striatal loops indicates the lack of downregulation of the subcortical areas. In **Figure 9**, TPDS—but not NBS—reported differences in the connectivity of the left hemisphere, amygdala, thalamus, and hippocampus, consistent with the valence hypothesis. However, the abundant presentation of right hemisphere ROIs in **Figure 9** supports the right hemisphere hypothesis indicating that the connectivity within the right hemisphere may be a biomarker for MDD. TPDS revealed a larger right hemisphere network which was sidestepped by NBS. This network is predominantly composed of the basal temporal lobe structures as well as occipital ROIs such as precuneus and pericalcarine. The difference in the temporal and parietal functionality in MDD is reported less in comparison with those in front striatal structures; however, there is a growing body of literature that focuses on the hypoactivity of the right hemisphere temporal areas in MDD (Bruder et al., 2017). The detection of such ROIs by TPDS is supportive of these studies reported in Bruder et al. (2017). Finally, several rsfMRI biomarkers of MDD are reported in Drysdale et al. (2017). After clustering these biomarkers through machine learning techniques, four different subtypes of MDD can be derived, based on four different clusters of ROIs. Unfortunately, the temporal areas of the brain are excluded in this study, due to a lack of data collection from several participating research sites. However, the ROI network reported by both NBS and TPDS in **Figure 9** is also reported in Drysdale et al. (2017), verifying our results in a much larger sample size.

In their meta-analysis of over 231 patients with MDD and 261 comparison participants, Yi Liao et al. found four consistent locations of decreased FA: white matter in the right frontal lobe, right fusiform gyrus, left frontal lobe, and right occipital lobe. Mainly, the right inferior longitudinal fasciculus, right inferior fronto-occipital fasciculus, and right posterior thalamic radiation were involved in such changes (Liao et al., 2013). This covers most of the connection pairs we have found in **Figure 9**, especially the right fusiform gyrus connections with R. Inferior temporal, parahipppocampal, and temporal gray matter are important because the NBS method failed to reveal all of these areas consistent with the meta-analysis.

In another meta-analysis (Wen et al., 2014), reduced FA is reported in the DLPFC and UF of patients with late-life depression (Wen et al., 2014). Those regions are part of frontostriatal and limbic networks consistent with our findings in **Figure 9**. This is also consistent with NBS analysis, especially the connections colored in green.

Another recent meta-analysis study has analyzed WM anisotropy and diffusivity in 1,305 MDD patients and 1,602 healthy controls (age range 12–88 years) from 20 samples worldwide (van Velzen et al., 2020). On adults, lower FA was observed in 16 of the 25 ROIs. The largest changes have been found mainly in the anterior corona radiata (ACR), corona radiata (CR), corpus callosum (CC), genu of the corpus callosum (GCC), body of the corpus callosum (BCC), and anterior limb of the internal capsule (ALIC). Significantly lower FA was also observed in the superior fronto-occipital fasciculus (SFO), sagittal stratum (SS), internal capsule (IC), posterior corona radiata (PCR), superior corona radiata (SCR), inferior fronto-occipital fasciculus (IFO), fornix/stria terminalis (FXST), external capsule (EC), and cingulate gyrus of the cingulum bundle (CGC). It is quite important to note that most of these regions are better fitted by TPDS in comparison with TBSS and VBA as revealed by our first test on these methods. The superior fronto-occipital fasciculus (left–right), sagittal stratum (left–right), superior corona radiata (left–right), posterior corona radiata (left–right), superior fronto-occipital fasciculus (left–right), inferior fronto-occipital fasciculus (left–right), external capsule (left–right), fornix (cres)/stria terminalis (left–right), and cingulum (left–right) are all better modeled using TPDS. This is also true for the anterior and superior corona radiata where only the right anterior corona radiata is modeled better with TBSS skeleton. The parts of the corpus callosum are on the other hand fitted better as the genu of the corpus callosum for TPDS, the body of the corpus callosum for VBA, and the splenium of the corpus callosum for TBSS. Overall, the benefit of TPDS is demonstrated in two different ways: 1. By fitting the underlying structural connections to an analytical model in a better way 2. By capturing wider network connectivity differences especially along longer tracts.

CONCLUSION

To conclude, we have shown that by analyzing PDDs using directional statistics, more insight is gained about fiber tracts regarding differences between populations. While other connectivity-based analysis methods may disregard the differences between longer fibers, TPDS becomes more robust as fiber tract length increases. In areas with low FA values, the distribution of PDDs among the fiber tracts can differentiate connectivity-based dysfunctions better, due to the power of directional statistics. The directional statistics analysis suggested here can also be applied by augmenting the existing methods, namely TBSS and VBA. Such an addition to the existing methods is valuable because it opens up the possibility to use parametric fitting along with directional statistics. The proposed method could be extended considering second and third directions of the diffusion tensor. In a future study, this can be modeled separately, fitting different distribution models for each direction

and analyzing the statistical changes of each direction in disease conditions.

When we implemented TPDS in two subject populations, one healthy and the other with MDD, we found several WM tract differences that are not reported in other methods such as NBS and TBSS. It is imperative to use TPDS on other subject populations and with more subjects to justify its strength in comparison with other methods that perform WM tract-based group analysis.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The data was collected from medication naive patients for a larger project. The entire project data will be released when the main research is published. Requests to access these datasets should be directed to DG, didemgokcay@gmail.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ankara university faculty of medicine clinical research ethics committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MM designed the model, computational framework, and carried out the implementation. DG directed the study on the MDD population and designed the data collection pipeline. Both authors analysed the data, discussed the results, and contributed to the final manuscript.

FUNDING

The authors appreciate METU (Middle East Technical University) and TUBITAK (Turkish National Science Foundation) for partially funding this project (project numbers BAP-07-04-2012 and 109E081).

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Prof. Bora Baskak from Ankara University, Ergin Atalay from Bilkent University, and Zeynep Başgöze for diagnosis of the MDD patients, data acquisition at the UMRAM center, and data collection. Special thanks to Gözde Ünal and Vilda Purutçuoglu for their comments and suggestions during the doctoral studies of MM through which the current study was accomplished.

REFERENCES

Alves, N. T., Fukusima, S. S., and Aznar-Casanova, J. A. (2008). Models of brain asymmetry in emotional processing. *Psychol. Neurosci.* 1, 63–66. doi: 10.3922/j.psns.2008.1.010

Basser, P. J. (1995). Inferring microstructural features and the physiological state of tissues from diffusion-weighted images. *NMR Biomed.* 8, 333–344.

Basser, P. J., and Pierpaoli, C. (2011). Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magn. Reson.* 213, 560–570. doi: 10.1016/j.jmr.2011.09.022

Behrman-Lay, A. M., Usher, C., Conturo, T. E., Correia, S., Laidlaw, D. H., Lane, E. M., et al. (2015). Fiber bundle length and cognition: a length-based tractography MRI study. *Brain Imaging Behav.* 9, 765–775. doi: 10.1007/s11682-014-9334-8

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *Ann. Stat.* 2, 1201–1225. doi: 10.1214/aos/1176342874

Blum, H. (1967). "A transformation for extracting new descriptors of shape," in *Models for the Perception of Speech and Visual Form*, Vol. 4. ed. W. Wathen-Dunn (MIT Press Cambridge: MIT Press), 153–171.

Bruder, G. E., Stewart, J. W., and McGrath, P. J. (2017). Right brain, left brain in depressive disorders: clinical and theoretical implications of behavioral, electrophysiological and neuroimaging findings. *Neurosci. Biobehav. Rev.* 78, 178–191. doi: 10.1016/j.neubiorev.2017.04.021

Castaneda, A. E., Tuulio-Henriksson, A., Marttunen, M., Suvisaari, J., and Lönnqvist, J. (2008). A review on cognitive impairments in depressive and anxiety disorders with a focus on young adults. *J. Affect. Disord.* 106, 1–27. doi: 10.1016/j.jad.2007.06.006

Cheng, Y., Xu, J., Yu, H., Nie, B., Li, N., Luo, C., et al. (2014). Delineation of early and later adult onset depression by diffusion tensor imaging. *PLoS One* 9:e0112307. doi: 10.1371/journal.pone.0112307

Cook, P. A., Bai, Y., Nedjati-Gilani, S., Seunarine, K. K., Hall, M. G., Parker, G. J., et al. (2005). Camino: open-source diffusion-MRI reconstruction and processing. *14th Scientific Meeting of the International Society for Magnetic Resonance in Medicine.* 2759–2760. http://www.cs.ucl.ac.uk/research/medic/camino

Corouge, I., Fletcher, P., Joshi, S., Gouttard, S., and Gerig, G. (2006). Fiber tract-oriented statistics for quantitative diffusion tensor MRI analysis. *Med. Image Anal.* 10, 786–798. doi: 10.1016/j.media.2006.07.003

Cullen, K. R., Klimes-Dougan, B., Muetzel, R., Mueller, B. A., Camchong, J., Houri, A., et al. (2010). Altered white matter microstructure in adolescents with major depression: a preliminary study. *J. Am. Acad. Child Adolesc. Psychiatry* 49, 173–183.e1.

Daducci, A., Gerhard, S., Griffa, A., Lemkaddem, A., Cammoun, L., Gigandet, X., et al. (2012). The connectome mapper: an open-source processing pipeline to map connectomes with MRI. *PLoS One* 7:e48121. doi: 10.1371/journal.pone.0048121

de Groot, M., Vernooij, M. W., Klein, S., Ikram, M. A., Vos, F. M., Smith, S. M., et al. (2013). Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. *NeuroImage* 76, 400–411. doi: 10.1016/j.neuroimage.2013.03.015

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Drevets, W. C., Price, J. L., and Furey, M. L. (2008). Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. *Brain Struct. Funct.* 213, 93–118. doi: 10.1007/s00429-008-0189-x

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* 23, 28–38. doi: 10.1038/nm.4246

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/S0896-6273(02)00569-X

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., et al. (2004). Automatically parcellating the human cerebral cortex. *Cereb. Cortex* 14, 11–22. doi: 10.1093/cercor/bhg087

Fisher, N. I., Lewis, T., and Embleton, B. J. J. (1993). *Statistical Analysis of Spherical Data*. Cambridge: Cambridge University Press.

Goodlett, C. B., Fletcher, P. T., Gilmore, J. H., and Gerig, G. (2009). Group analysis of DTI fiber tract statistics with application to neurodevelopment. *NeuroImage* 45, S133–S142. doi: 10.1016/j.neuroimage.2008.10.060

Hecke, W. V., Leemans, A., de Backer, S., Jeurissen, B., Parizel, P. M., and Sijbers, J. (2009). Comparing isotropic and anisotropic smoothing for voxel-based DTI analyses: a simulation study. *Hum. Brain Mapp.* 31, 98–114. doi: 10.1002/hbm.20848

Helm, K., Viol, K., Weiger, T. M., Tass, P. A., Grefkes, C., del Monte, D., et al. (2018). Neuronal connectivity in major depressive disorder: a systematic review. *Neuropsychiatr. Dis. Treat.* 14, 2715–2737. doi: 10.2147/NDT.S170989

Hutchinson, E. B., Rutecki, P. A., Alexander, A. L., and Sutula, T. P. (2012). Fisher statistics for analysis of diffusion tensor directional information. *J. Neurosci. Methods* 206, 40–45. doi: 10.1016/j.jneumeth.2012.02.004

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). Review FSL. *NeuroImage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/S1361-8415(01)00036-6

Jones, D., Symms, M., Cercignani, M., and Howard, R. (2005). The effect of filter size on VBM analyses of DT-MRI data. *Neuroimage* 26, 546–554. doi: 10.1016/j.neuroimage.2005.02.013

Jones, D. K., and Cercignani, M. (2010). Twenty-five pitfalls in the analysis of diffusion MRI data. *NMR Biomed.* 23, 803–820. doi: 10.1002/nbm.1543

Kieseppä, T., Eerola, M., Mäntylä, R., Neuvonen, T., Poutanen, V. P., Luoma, K., et al. (2010). Major depressive disorder and white matter abnormalities: a diffusion tensor imaging study with tract-based spatial statistics. *J. Affect. Disord.* 120, 240–244. doi: 10.1016/j.jad.2009.04.023

Kong, T. Y., and Rosenfeld, A. (1989). Digital topology: introduction and survey. *Comput. Vis. Graph. Image Process.* 48, 357–393. doi: 10.1016/0734-189X(89)90147-3

Liao, Y., Huang, X., Wu, Q., Yang, C., Kuang, W., Du, M., et al. (2013). Is Depression a disconnection syndrome? Meta- analysis of diffusion tensor imaging studies in patients with MDD. *J. Psychiatry Neurosci.* 38, 49–56. doi: 10.1503/jpn.110180

Mandl, R. W., Schnack, H. G., Luigjes, J., van den Heuvel, M. P., Cahn, W., Kahn, R. S., et al. (2010). Tract-based analysis of magnetization transfer ratio and diffusion tensor imaging of the frontal and frontotemporal connections in schizophrenia. *Schizophrenia Bull.* 36, 778–787. doi: 10.1093/schbul/sbn161

Mardia, K. V., and Jupp, PE (eds) (1999). "Directional statistics," in *Wiley Series in Probability and Statistics*, (Hoboken, NJ: John Wiley & Sons, Inc), doi: 10.1002/9780470316979

Mardia, K. V., and Zemroch, P. J. (1977). Table of maximum likelihood estimates for the bingham distribution. *J. Stat. Comput. Simul.* 6, 29–34. doi: 10.1080/00949657708810165

Mayberg, H. S. (2003). Modulating dysfunctional limbic-cortical circuits in depression: towards development of brain-based algorithms for diagnosis and optimised treatment. *Br. Med. Bull.* 65, 193–207. doi: 10.1093/bmb/65.1.193

McClintock, S. M., Husain, M. M., Greer, T. L., and Cullum, C. M. (2010). Association between depression severity and neurocognitive function in major depressive disorder: a review and synthesis. *Neuropsychology* 24, 9–34. doi: 10.1037/a0017336

Metin, M. O., and Gökçay, D. (2014). Direction statistics methods evaluation for principal diffusion directions. *Hum. Brain Mapp.*

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., et al. (2008). Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. *NeuroImage* 40, 570–582. doi: 10.1016/j.neuroimage.2007.12.035

Pajevic, S., and Pierpaoli, C. (1999). Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data. *Magn. Reson. Med.* 42, 526–540.

Palágyi, K., Balogh, E., Kuba, A., Halmai, C., Erdőhelyi, B., Sorantin, E., et al. (2001). "A sequential 3D thinning algorithm and its medical applications," in *Information Processing in Medical Imaging. IPMI 2001. Lecture Notes in Computer Science*, eds M. F. Insana and R. M. Leahy (Berlin: Springer), 409–415. doi: 10.1007/3-540-45729-1_42

Parker, G. J. M., Haroon, H. A., and Wheeler-Kingshott, C. A. M. (2003). A framework for a streamline-based probabilistic index of connectivity (PICo) using a structural interpretation of MRI diffusion measurements. *J. Magn. Reson. Imag.* 18, 242–254. doi: 10.1002/jmri.10350

Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: basic tools for geometric measurements. *J. Math. Imaging Vis.* 25, 127–154. doi: 10.1007/s10851-006-6228-4

Riffert, T. W., Schreiber, J., Anwander, A., and Knösche, T. R. (2014). Beyond fractional anisotropy: extraction of bundle-specific structural metrics from crossing fiber models. *NeuroImage* 100, 176–191. doi: 10.1016/j.neuroimage.2014.06.015

Schwartzman, A., Dougherty, R. F., and Taylor, J. E. (2005). Cross-subject comparison of principal diffusion direction maps. *Magn. Reson. Med.* 53, 1423–1431. doi: 10.1002/mrm.20503

Seminowicz, D. A., Mayberg, H. S., McIntosh, A. R., Goldapple, K., Kennedy, S., Segal, Z., et al. (2004). Limbic–frontal circuitry in major depression: a path modeling metanalysis. *NeuroImage* 22, 409–418. doi: 10.1016/j.neuroimage.2004.01.015

Sexton, C. E., Mackay, C. E., and Ebmeier, K. P. (2009). A systematic review of diffusion tensor imaging studies in affective disorders. *Biol. Psychiatry* 66, 814–823. doi: 10.1016/j.biopsych.2009.05.024

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31, 1487–1505. doi: 10.1016/j.neuroimage.2006.02.024

Song, S. K., Sun, S. W., Ramsbottom, M. J., Chang, C., Russell, J., and Cross, A. H. (2002). Dysmyelination revealed through MRI as increased radial (but Unchanged Axial) diffusion of water. *NeuroImage* 17, 1429–1436.

Sotiropoulos, S. N., Behrens, T. E. J., and Jbabdi, S. (2012). Ball and rackets: inferring fiber fanning from diffusion-weighted MRI. *NeuroImage* 60, 1412–1425. doi: 10.1016/j.neuroimage.2012.01.056

Taoka, T., Aida, N., Fujii, Y., Ichikawa, K., Kawai, H., Nakane, T., et al. (2020). White matter microstructural changes in tuberous sclerosis: evaluation by neurite orientation dispersion and density imaging (NODDI) and diffusion tensor images. *Sci. Rep.* 10, 1–9. doi: 10.1038/s41598-019-57306-w

Tariq, M., Schneider, T., Alexander, D. C., Wheeler-Kingshott, C. A. G., and Zhang, H. (2016). Bingham-NODDI: mapping anisotropic orientation dispersion of neurites using diffusion MRI. *NeuroImage* 133, 207–223. doi: 10.1016/j.neuroimage.2016.01.046

Timmers, I., Roebroeck, A., Bastiani, M., Jansma, B., Rubio-Gozalbo, E., and Zhang, H. (2016). Assessing microstructural substrates of white matter abnormalities: a comparative study using DTI and NODDI. *PLoS One* 11:e167884. doi: 10.1371/journal.pone.0167884

van Velzen, L. S., Kelly, S., Isaev, D., Aleman, A., Aftanas, L. I., Bauer, J., et al. (2020). White matter disturbances in major depressive disorder: a coordinated analysis across 20 International cohorts in the ENIGMA MDD Working Group. *Mol. Psychiatry* 25, 1511–1525. doi: 10.1038/s41380-019-0477-2

Wang, R., Benner, T., Sorensen, A., and Wedeen, V. J. (2007). Diffusion toolkit: a software package for diffusion imaging data processing and tractography. *Proc Int Soc Mag Reson Med* 15.

Wassermann, D., Rathi, Y., Bouix, S., Kubicki, M., Kikinis, R., Shenton, M., et al. (2011). "White matter bundle registration and population analysis based on gaussian processes," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6801 LNCS*, (Stapleton, NY: NIH Public Access), 320–332. doi: 10.1007/978-3-642-22092-0_27

Watson, G. S., and Williams, E. J. (1956). On the construction of significance tests on the circle and the sphere. *Biometrika* 43:344. doi: 10.2307/2332913

Wen, M. C., Steffens, D. C., Chen, M. K., and Zainal, N. H. (2014). Diffusion tensor imaging studies in late-life depression: systematic review and meta-analysis. *Int. J. Geriatr. Psychiatry* 29, 1173–1184. doi: 10.1002/gps.4129

Zalesky, A., Fornito, A., and Bullmore, E. T. (2010). Network-based statistic: identifying differences in brain networks. *NeuroImage* 53, 1197–1207. doi: 10.1016/j.neuroimage.2010.06.041

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage* 61, 1000–1016. doi: 10.1016/j.neuroimage.2012.03.072

Zou, K., Huang, X., Li, T., Gong, Q., Li, Z., Ou-Yang, L., et al. (2008). Alterations of white matter integrity in adults with major depressive disorder: a magnetic resonance imaging study. *J. Psychiatry Neurosci.* 33, 525–530.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 06 April 2021
doi: 10.3389/fnins.2021.647535

GAMER-MRI in Multiple Sclerosis Identifies the Diffusion-Based Microstructural Measures That Are Most Sensitive to Focal Damage: A Deep-Learning-Based Analysis and Clinico-Biological Validation

*Po-Jui Lu[1,2,3], Muhamed Barakovic[1,2,3], Matthias Weigel[1,2,3,4], Reza Rahmanzadeh[1,2,3], Riccardo Galbusera[1,2,3], Simona Schiavi[5], Alessandro Daducci[5], Francesco La Rosa[6,7,8], Meritxell Bach Cuadra[6,7,8], Robin Sandkühler[9], Jens Kuhle[2,3], Ludwig Kappos[2,3], Philippe Cattin[9] and Cristina Granziera[1,2,3]***

[1] Translational Imaging in Neurology (ThINk) Basel, Department of Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland, [2] Neurologic Clinic and Policlinic, Departments of Medicine, Clinical Research and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland, [3] Research Center for Clinical Neuroimmunology and Neuroscience Basel, University Hospital Basel and University of Basel, Basel, Switzerland, [4] Division of Radiological Physics, Department of Radiology, University Hospital Basel, Basel, Switzerland, [5] Department of Computer Science, University of Verona, Verona, Italy, [6] Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [7] CIBM Center for Biomedical Imaging, Lausanne, Switzerland, [8] Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland, [9] Center for Medical Image Analysis and Navigation, Department of Biomedical Engineering, Faculty of Medicine, University of Basel, Allschwil, Switzerland

Conventional magnetic resonance imaging (cMRI) in multiple sclerosis (MS) patients provides measures of focal brain damage and activity, which are fundamental for disease diagnosis, prognosis, and the evaluation of response to therapy. However, cMRI is insensitive to the damage to the microenvironment of the brain tissue and the heterogeneity of MS lesions. In contrast, the damaged tissue can be characterized by mathematical models on multishell diffusion imaging data, which measure different compartmental water diffusion. In this work, we obtained 12 diffusion measures from eight diffusion models, and we applied a deep-learning attention-based convolutional neural network (CNN) (GAMER-MRI) to select the most discriminating measures in the classification of MS lesions and the perilesional tissue by attention weights. Furthermore, we provided clinical and biological validation of the chosen metrics—and of their most discriminative combinations—by correlating their respective mean values in MS patients with the corresponding Expanded Disability Status Scale (EDSS) and the serum level of neurofilament light chain (sNfL), which are measures of disability and neuroaxonal damage. Our results show that the neurite density index from neurite orientation and dispersion density imaging (NODDI), the measures of the intra-axonal and isotropic compartments from microstructural Bayesian approach, and the measure of the intra-axonal compartment from the spherical mean technique NODDI were the

OPEN ACCESS

Edited by:
Yu-Chien Wu, Indiana University Bloomington, United States

Reviewed by:
Rajikha Raja, Georgia State University, United States
Lipeng Ning, Brigham and Women's Hospital and Harvard Medical School, United States

***Correspondence:**
Cristina Granziera
cristina.granziera@unibas.ch

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *30 December 2020*
Accepted: *23 February 2021*
Published: *06 April 2021*

Citation:
Lu P-J, Barakovic M, Weigel M, Rahmanzadeh R, Galbusera R, Schiavi S, Daducci A, La Rosa F, Bach Cuadra M, Sandkühler R, Kuhle J, Kappos L, Cattin P and Granziera C (2021) GAMER-MRI in Multiple Sclerosis Identifies the Diffusion-Based Microstructural Measures That Are Most Sensitive to Focal Damage: A Deep-Learning-Based Analysis and Clinico-Biological Validation. Front. Neurosci. 15:647535. doi: 10.3389/fnins.2021.647535

most discriminating (respective attention weights were 0.12, 0.12, 0.15, and 0.13). In addition, the combination of the neurite density index from NODDI and the measures for the intra-axonal and isotropic compartments from the microstructural Bayesian approach exhibited a stronger correlation with EDSS and sNfL than the individual measures. This work demonstrates that the proposed method might be useful to select the microstructural measures that are most discriminative of focal tissue damage and that may also be combined to a unique contrast to achieve stronger correlations to clinical disability and neuroaxonal damage.

Keywords: multiple sclerosis, deep learning, advanced quantitative diffusion MRI, relative importance order, clinically correlated measure selection

INTRODUCTION

Conventional magnetic resonance imaging (cMRI) in multiple sclerosis (MS) plays a major role in MS diagnosis, prognosis, and in the evaluation of patients' therapeutic response (Rovira et al., 2015; Wattjes et al., 2015). However, the heterogeneity of focal MS lesions, the pathology in normal-appearing white and gray matter (NAWM and NAGM), and the specific damage to myelin and axons are largely overlooked by cMRI. Multishell diffusion-weighted imaging (mDWI) provides a way to further probe tissue damage and repair in MS patients (Schneider et al., 2017; Lakhani et al., 2020). mDWI measures signal changes that are related to the diffusion of water molecules within central nervous system (CNS) tissue (Novikov et al., 2019; Lakhani et al., 2020), which is constrained by the local microenvironment (Novikov et al., 2019). This enables diffusion measures of biophysical microstructure models derived from mDWI to decode the information specific to different water compartments (e.g., intra-axonal and isotropic compartments) within the CNS tissue (Novikov et al., 2019). The intra-axonal compartment reflects the integrity of the neurites, and the isotropic compartment indicates the movement of the free water (Novikov et al., 2019). These two compartments can describe the two pathological presentations of MS lesions, demyelination, and axonal injury and are commonly modeled by various biophysical microstructure models (Lakhani et al., 2020).

A microenvironment characteristic is measured differently by the measures from different mathematical models due to the different assumptions on the diffusion within the tissue. Yet, to our knowledge, the direct comparison of all considered diffusion measures on MS lesions and the possibility to combine them does not exist. Therefore, how to select the most discriminating diffusion measures for a given neurological disorder and how to combine the complementary information they might provide remain to be open questions and motivate this study.

Convolutional neural network (CNN) in deep learning has proven to be promising in various applications of MR images and is able to encode spatial patterns on the images into representative hidden features (Andermatt et al., 2018; Yoo et al., 2018; Akçakaya et al., 2019; La Rosa et al., 2020; Saha et al., 2020). In our previous work (Lu et al., 2020), we used an attention-based CNN—GAMER-MRI—to rank the importance of the input quantitative MRIs in the classification of stroke and MS lesions. Here, we further developed the method to select discriminating intercorrelated diffusion measures in the classification of MS lesions and the perilesional tissue. Compared to the conventional feature selection methods, this CNN-based method enables utilizing maximally available spatial information of the images and does not need to decide on how to find representative values for the samples of each contrasts, such as the mean value only within a lesion neglecting the perilesion tissue. In addition, the method jointly considers all the contrasts, which is a limitation for most of the conventional feature selection methods. Furthermore, in this study, we have explored the relationship between the chosen measures, or their combinations, with the Expanded Disability Status Scale (EDSS) and the neurofilament light chain in the serum (sNfL), which are respectively (i) a clinical measure of disability in MS patients and (ii) a biological measure of neuroaxonal damage (Barro et al., 2018; Siller et al., 2019).

MATERIALS AND METHODS

MRI Data

One hundred twenty-three MS patients (84 relapsing–remitting and 39 progressive, 71 female and 52 male, age range = 44.7 ± 14.0, median EDSS = 2.5, EDSS range of 0.0–8.0) were enrolled in the study, which was approved by the local Ethics Committee of Basel University Hospital. All subjects gave written consent prior to the enrollment. MS patients underwent a multiparametric protocol on 3T whole-body MR system (Siemens MAGNETOM Prisma). The protocol included 3D SPACE-based FLAIR, 3D magnetization-prepared 2 rapid gradient echoes (MP2RAGE) (Marques et al., 2010), and mDWI (**Table 1**).

Measured diffusion-weighted imaging was denoised by MRtrix (Cordero-Grande et al., 2019; Tournier et al., 2019). The correction of susceptibility-induced distortion with the reversed phase-encoding images, eddy currents, and movement was performed by FMRIB Software Library (FSL) (Andersson et al., 2003; Smith et al., 2004; Jenkinson et al., 2012; Andersson and Sotiropoulos, 2016). The quantitative diffusion measures for the isotropic and intra-axonal compartments were reconstructed from the eight open-source biophysical

TABLE 1 | Acquisition parameters of each contrast in the MS dataset.

	TE (ms)	TR (ms)	FOV (mm^3)	SR (mm^3)	TI (ms)	Additional parameters
FLAIR	386	5000	256 × 256 × 256	1 × 1 × 1	1800	–
MP2RAGE	3	5000	256 × 256 × 256	1 × 1 × 1	700, 2500	–
						b values (s/mm^2)
mDWI	75	4500	256 × 256 × 144	1.8 × 1.8 × 1.8	–	0/12 acquisitions and 12 reverse encoding acquisitions; 700; 1000; 2000; 3000/137 directions in total

TE, echo time; TR, repetition time; TI, inversion time; FOV, field of view; SR, spatial resolution.

models, including Ball and Stick[1] (Behrens et al., 2003), neurite orientation and dispersion density imaging (NODDI)[2] (Zhang et al., 2012), NODDI with the spherical mean technique (SMT-NODDI)[1] (Cabeen et al., 2019), microstructure Bayesian (MB) approach[3] (Reisert et al., 2017), multicompartment microscopic diffusion imaging (MCMDI)[1] (Kaden et al., 2016), neurite orientation dispersion and density imaging with diffusivities assessment (NODDIDA)[4] (Jelescu et al., 2015), distribution of 3D anisotropic microstructural environments in diffusion-compartment imaging (DIAMOND)[5] (Scherrer et al., 2016), and microstructure fingerprinting[6] (Rensonnet et al., 2019). The exemplary diffusion measures and FLAIR are in **Figure 1**.

The quantitative diffusion measures of each patient were masked by the brain mask to remove non-brain tissue including the ventricle. The brain mask was the binarized subcortical segmentation obtained from FreeSurfer (Fischl et al., 2001) on MP2RAGE (Fujimoto et al., 2014) and transformed by FSL to align with mDWI. The diffusion measures were then subject-wise normalized. Eighty-four patients were randomly selected to be used in a 5-fold cross-validation. The other 39/123 patients formed a pure test dataset. White matter lesions were automatically segmented using FLAIR and MP2RAGE[7] (La Rosa et al., 2020) and manually corrected by two expert raters. The lesion segmentations were transformed by FSL to be aligned with mDWI. Lesions of size less than three voxels were excluded. The perilesional tissue was defined as white matter tissue locating within a three-voxel region around the lesions. Patches of 5 × 5 × 5 voxels were sampled on lesions and perilesional tissue considering the lesion sizes. To reduce the overlapping between the lesion and perilesional patches due to their proximity, a constraint of at most 20% of a sampled patch being overlapped with another patch was applied. The numbers of patches being sampled on each lesion and perilesional tissue were proportional to the size of the lesion and the perilesional tissue, respectively. In the end, 3007 lesion patches and 3624 perilesional patches were sampled in the dataset for 5-fold cross-validation, and 1402 lesion patches and 1665 perilesional patches were sampled in the pure test dataset. The 5-fold cross-validation was based on the number of patients. Therefore, patches from a patient would not present both in the training and in the validation datasets.

GAMER-MRI

GAMER-MRI was previously developed and validated as a method to obtain attention weights and the relative importance in a classification task of given input contrasts (Lu et al., 2020). As we previously reported, the neural network consisted of three parts for feature extraction, gated attention mechanism (Ilse et al., 2018), and classification (Lu et al., 2020). The feature extraction part included three convolutional blocks for each contrast. Each convolutional block was composed of a layer of 16 convolutional filters and exponential leaky units followed by batch normalization. The kernel size of the convolutional filter was 3 × 3 × 3, and padding was applied correspondingly to maintain the patch size. After the last convolutional block, a 16-neuron fully connected layer (FCL) received the flattened vector of 125 elements and encoded the hidden feature of 16 elements. The gated attention mechanism was formed by an attention layer containing an eight-neuron FCL followed by the tanh function and a gate layer having an eight-neuron FCL followed by the sigmoid function. The outputs of tanh and sigmoid were element-wise multiplied. From the element-wise product, in the original implementation for not-highly-correlated input contrasts, the attention weights were obtained by following one-neuron FCL and the softmax function (Lu et al., 2020). However, this design was not effective for highly correlated inputs, i.e., diffusion measures in this work. The information content of measures is similar, and thus, the difference in the obtained attention weights was small.

For the purpose of this study, we multiplied the outputs from the element-wise multiplication by 2. This enhanced the difference between the encoded features of the correlated diffusion measures during training because the exponential transformation in the softmax function could not properly reflect the difference in the small and negative values. For example, 0.01 is 10 times larger than 0.001, but they become 1.01 and 1.001 after the exponential transformation. This leads to 0.502 and 0.498 as attention weights after the softmax function. The enhanced output was then connected to a one-neuron FCL followed by the softmax function to generate the normalized attention weights. The weighted sum of the hidden features and the corresponding attention weights formed a combined hidden feature for the classifier. The classifier was one sigmoid neuron. The network structure is in **Figure 2**.

[1]https://github.com/AthenaEPI/dmipy
[2]https://github.com/daducci/AMICO
[3]https://bitbucket.org/reisert/baydiff/wiki/Home
[4]https://github.com/robbert-harms/MDT
[5]https://bitbucket.org/benoitscherrer/crldciestimate
[6]The author needs to be contacted.
[7]https://hub.docker.com/r/francescolr/ms_seg

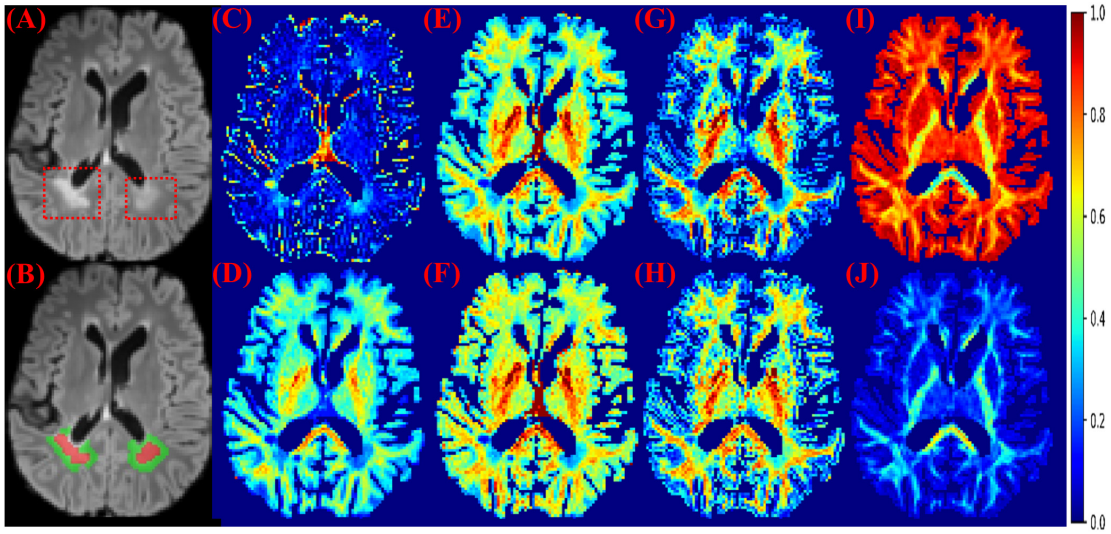


FIGURE 1 | MS lesions on FLAIR and diffusion measures. **(A)** FLAIR: MS lesions are hyperintense and indicated by red dashed boxes. **(B)** Red: lesions; Green: perilesional white matter tissue. **(C)** The isotropic compartment from MB. **(D)** The intra-axonal compartment from MB. **(E)** The neurite density index from NODDI. **(F)** The intra-axonal compartment from SMT-NODDI. **(G)** The intra-axonal compartment from MCMDI. **(H)** The intra-axonal compartment from NODDIDA. **(I)** The isotropic compartment from Ball and Stick. **(J)** The intra-axonal compartment from Ball and Stick. Other measures in the analysis are in **Supplementary Figure 1**.

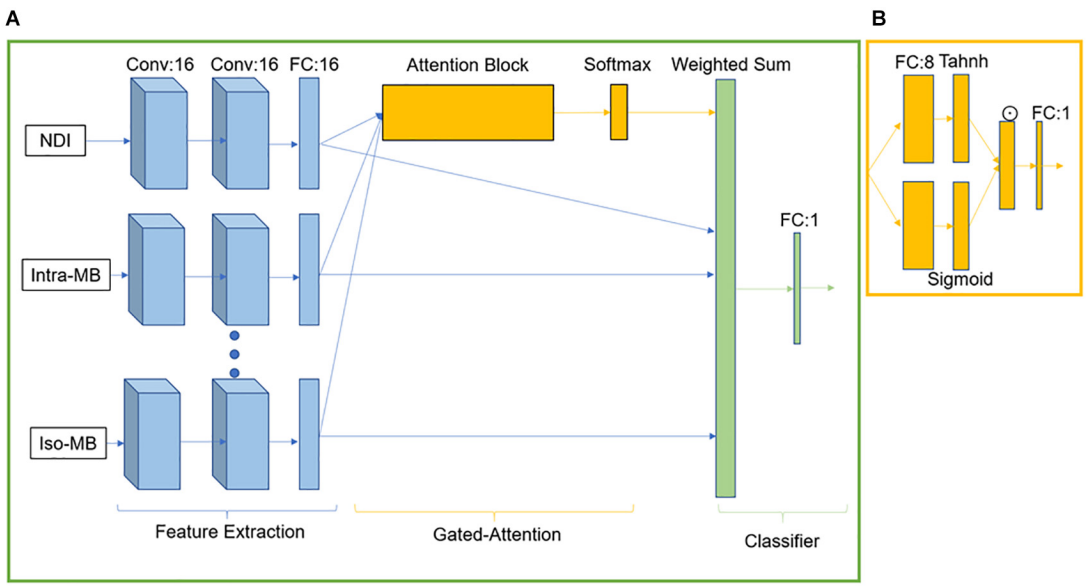


FIGURE 2 | GAMER-MRI. **(A)** The neural network. Conv stands for the convolutional block. FC is a fully connected layer. **(B)** Attention block. ⊙ represents an element-wise multiplication.

The weighted sampler was used to account for the class imbalance, and the batch size was 256. The loss function was cross-entropy loss. The evaluation metric was the area under the receiver operating characteristic curve (AUC). The optimizer was AdamW (Loshchilov and Hutter, 2019) with the learning rate = 5e-5 and the weight decay = 1e-2. To avoid overfitting, data augmentation and a learning-rate scheduler were performed. On-the-fly data augmentation included random flipping in the left–right directions and Gaussian noise with zero mean and unit standard deviation. The scheduler was the learning-rate-reduce-plateau scheduler with a patience of 15 epochs.

Selection of Contrasts

Intrinsic strong correlation between the quantitative diffusion measures can lead to instability of the obtained attention weights and the ranked order, compared to the result in Lu et al. (2020). Therefore, to avoid determination solely based on the attention weights, the selection of discriminating measures

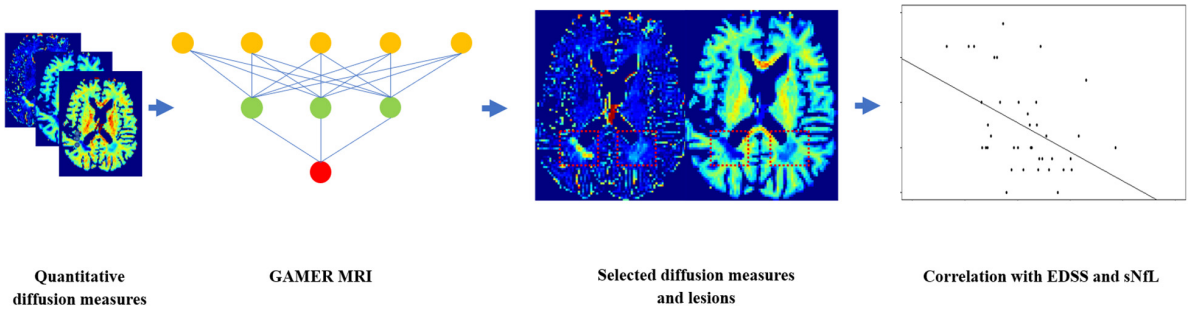


FIGURE 3 | Flowchart for using GAMER-MRI to select the most discriminating subject-wise normalized diffusion measures and correlating the combinations of selected diffusion measures with the Expanded Disability Status Scale and the serum level of neurofilament light chain.

was an iteration process. It started from the measure whose attention weight was dominant in the validation datasets in all the cross-validation folds. If no measure was selected, the measures whose attention weights were ranked first or second in all the folds were considered. If no measures were selected, the attention weights that ranked first or second and third in all the folds were considered. The selection stopped when the sum of their attention weights was over 0.5, which meant that the selected measures were more important than 50% of the input diffusion measures in differentiating the lesion and perilesional tissue.

To assess which selected subject-wise normalized quantitative diffusion measures, or combination of those measures, was best correlated with patients' EDSS as well as NfL in the pure test dataset, we first averaged the diffusion measures within each lesion and then over lesions within each patient. In 31/39 patients of the test dataset, we quantified sNfL. Then, we performed Spearman's correlation coefficient with two-sided 20,000 permutation tests. The Benjamin–Hochberg procedure (Benjamini and Hochberg, 1995) was performed to control the false discovery rate (FDR) with the threshold 0.05. The flowchart is shown in **Figure 3**.

RESULTS

Lesion Classification

In **Table 2**, we report the average performance of GAMER-MRI using all the diffusion measures on the (i) validation dataset over fivefold cross-validation and (ii) on the pure test dataset.

The diffusion measures selected by using the validation datasets were the neurite density index (NDI) from NODDI, the intra-axonal and isotropic compartment from MB (Intra-MB and Iso-MB), and the intra-axonal compartment from SMT-NODDI (Intra-SMT) in **Figure 1**. Their average attention weights of the corrected predicted samples are also reported in **Table 2**.

Spearman's Correlation

Correlation With EDSS

The Spearman's correlation coefficients (ρ) and the corresponding original p-values of the selected normalized diffusion measures, or their statistically significant combinations and EDSS, are reported in **Table 3**. The Spearman's correlation coefficients (ρ) of the conventional lesion load metrics are also reported. The number of potential combinations of four selected diffusion measures is 15, and there are two tests in the lesion load analysis. This led to in total 17 statistical tests. The significance controlled by FDR is indicated by an asterisk. The scatter plot of the combination having the strongest correlation is in **Figure 4A**, and an exemplary image of the combination is in **Figure 4B**.

Correlation With sNfL

The Spearman's correlation coefficients (ρ) and the corresponding original p-values are reported in **Table 4**. One patient had a relatively high sNfL level of 160 μg/ml, compared to the mean sNfL level of 8.9 μg/ml of the rest of 30 patients. After this patient's data were excluded, the significance in **Table 4** did not change, but the correlation was stronger. For illustration purpose, the scatter plot of the combination

TABLE 2 | Performance of the patch-based network on MS lesions and the selected diffusion measures on fivefold cross-validation (first row, average mean, and standard deviation are reported) and pure testing set (second row). Balanced accuracy is defined as the average of sensitivity and specificity in each fold. FI score is defined as the harmonic mean of precision and recall.

Mean metrics (%)	**AUC**	**Balanced accuracy**	**Sensitivity**	**Specificity**	**F1 score**
Validation dataset	90.67 ± 0.009	83.26 ± 1.35	81.09 ± 2.44	85.44 ± 2.03	81.62 ± 1.67
Test dataset	91.01 ± 0.003	83.42 ± 0.12	83.39 ± 0.67	83.45 ± 0.82	82.14 ± 0.11
Selected measures	**NDI**	**Intra-MB**	**Iso-MB**	**Intra-SMT**	
Attention weights	0.121 ± 0.014	0.117 ± 0.014	0.145 ± 0.007	0.131 ± 0.015	

TABLE 3 | Spearman's correlation of selected normalized diffusion measures, or their combinations and EDSS.

Lesion load	ρ	*P*-value	Significance
Number of lesions	0.13	0.41	–
Lesion volume	0.25	0.12	–
Normalized diffusion measures			
NDI	−0.38	0.017	*
Intra-SMT	−0.31	0.057	
Intra-MB	−0.40	0.013	*
Iso-MB	0.09	0.58	–
Intra-MB + Iso-MB	−0.39	0.014	*
Intra-MB + NDI	−0.43	0.007	*
Intra-SMT + NDI	−0.37	0.023	*
Intra-SMT + Intra-MB	−0.40	0.012	*
Intra-MB + Iso-MB + NDI	−0.45	0.004	*
Intra-MB + Iso-MB + Intra-SMT	−0.42	0.007	*
Intra-MB + Intra-SMT + NDI	−0.42	0.009	*
Intra-MB + Iso-MB + NDI + Intra-SMT	−0.41	0.009	*

The significance is controlled by FDR with a threshold of 0.05. Only the combinations of significance are reported.

having the strongest correlation (**Figure 5A**) does not contain this outlier patient. An exemplary image of the combination is in **Figure 5B**.

DISCUSSION

Our work provided evidence that a modified version of GAMER-MRI, including a specific selection procedure for correlated measures, permits to identify the microstructural diffusion measures that are most discriminative of focal MS pathology among the ones obtained with eight open-source mathematical models of multishell diffusion data. Moreover, our data showed that some of the combinations of the selected normalized diffusion measures better correlated with patients' disability and neuroaxonal damage than the individual measures.

Diffusion-based microstructural measures quantify different compartments based on various assumptions. Nevertheless, the relative sensitivity of the different diffusion-based microstructural metrics to specific CNS pathologies is unclear. In this work, we have provided a methodological frame to discriminate the most sensitive diffusion microstructural measures to focal MS pathology in a large population of MS patients.

We first aimed at identifying which measure best discriminated MS lesions from the perilesional tissue because

TABLE 4 | Spearman's correlation of selected normalized diffusion measures, or their combinations and sNfL.

Lesion load	ρ	*P*-value	Significance
Number of lesions	0.48	0.006	*
Lesion volume	0.45	0.01	*
Normalized diffusion measures			
NDI	−0.37	0.04	–
Intra-SMT	−0.27	0.14	–
Intra-MB	−0.42	0.02	*
Iso-MB	0.1	0.59	–
Intra-MB + Iso-MB	−0.51	0.004	*
Intra-MB + NDI	−0.43	0.02	*
Intra-MB + Iso-MB + NDI	−0.48	0.007	*
Intra-MB + Iso-MB + Intra-SMT	−0.45	0.01	*
Intra-MB + Iso-MB + NDI + Intra-SMT	−0.44	0.02	*

The significance controlled by FDR with a threshold of 0.05. Only the combinations of significance are reported.

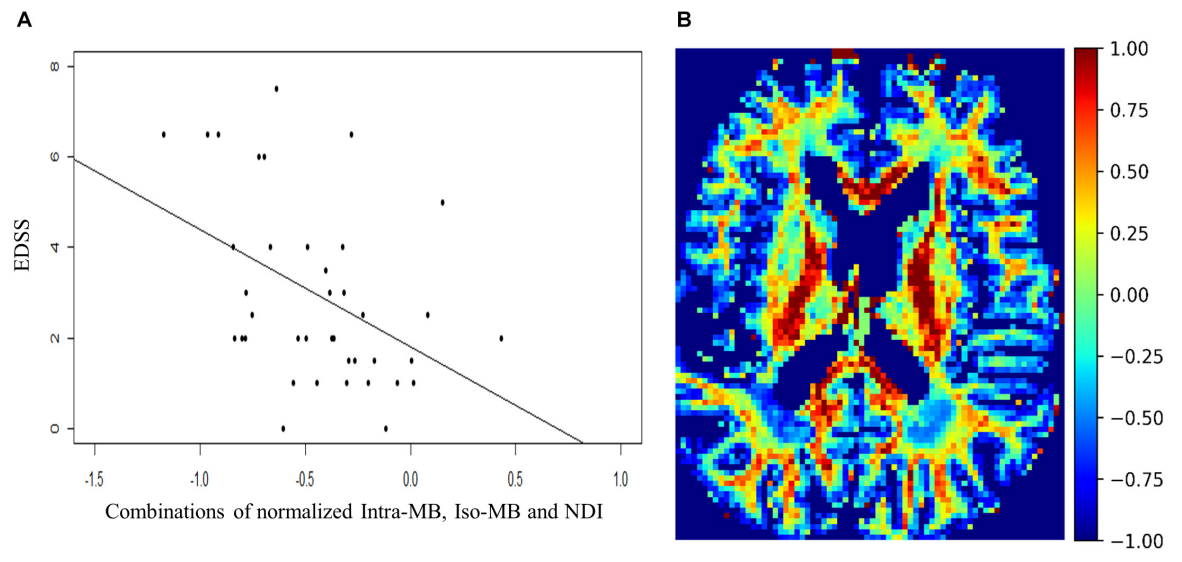


FIGURE 4 | (A) Scatter plot and a regression line of EDSS and the combinations of normalized Intra-MB, Iso-MB, and NDI, which has strongest correlation. **(B)** An exemplary image of the combined contrast.

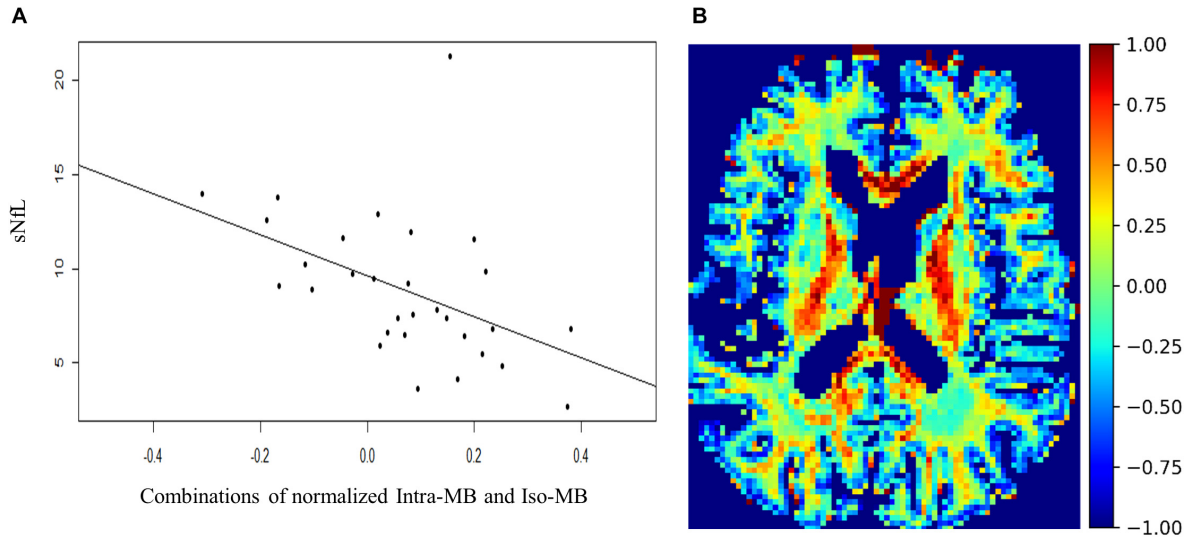


FIGURE 5 | (A) Scatter plot and a regression line of the combinations of normalized Intra-MB and Iso-MB, which showed the strongest correlation with sNFL. **(B)** An exemplary image of the combined contrast.

we judged that if the neural network was able to differentiate between lesions and the immediate surrounding tissue, the learned pattern would have been most sensitive to focal MS pathology than the one we would have derived by comparing lesions to the distant normal-appearing tissue. The evaluation metrics in **Table 2** indicated that the neural network was able to learn pivotal information for the target classification. As expected, because of the highly correlated nature of the studied diffusion-based measures, the difference among the obtained attention weights was small. The proposed selection process alleviated the fluctuating order of attention weights due to their small differences. The threshold of 0.5 in the selection process was empirically chosen considering the representativeness of selected diffusion measures and the multiple comparison problem.

The core idea of the attention mechanism is to enhance important features from the data themselves relevant to the specific application (Bahdanau et al., 2015). Therefore, in most of the applications in natural language processing and natural image classification, the attention weights were used to enhance the connection to the corresponding features based on their importance instead of quantifying the relative importance among the features (Maicas et al., 2017; Vaswani et al., 2017; Hu et al., 2018; Woo et al., 2018). Using different designs of the attention mechanism, the attention weights also provide the relative importance among features as shown in a histopathological image classification and image captioning (You et al., 2016; Ilse et al., 2018). In GAMER-MRI, attention weights were computed and validated on multicontrast MRI measures in order to select their relative importance in a given neurological disease classification.

To our knowledge so far, only few studies applied measures derived from microstructural models to study focal MS pathology (for a review, see Granziera et al., 2020) and only one study used deep-learning to show the superior performance of diffusion basis spectrum imaging to segment voxel-wise different types of MS lesions compared to using diffusion tensor imaging (Ye et al., 2020). However, the joint comparison of multiple microstructural diffusion measures in MS lesions has not been explored yet. This work considered the potential interaction between the measures and tried to address this issue.

The four selected diffusion measures include three measures for the intra-axonal compartment from three models and one measure for the isotropic compartment from one of the three models. This means that most of the discriminating information of the damaged neurons was from the loss of axonal integrity. The additional information about the inflammatory processes might be reflected by the measure for the isotropic compartment to better characterize the distinction of lesions.

Besides, by combining the selected diffusion measures in the discrimination of focal pathology, it was possible to achieve a stronger correlation with patient disability than one of those metrics alone or even conventional MRI metrics, such as the lesion number and volume. These results suggest that a comprehensive description of the tissue microstructure in regions of focal damage in MS patients may well help decrease the clinical–radiological paradox (Barkhof, 2002). Interestingly, the combined contrast achieving the best correlation with disability was the sum of measures quantifying intra-axonal and isotropic diffusion, which may be considered surrogate measures of the loss of integrity of axons and myelin as well as of inflammatory processes (i.e., increased cellularity and edema).

Most of the combinations that best correlated to EDSS were also highly related to the sNfL levels: remarkably, the correlation coefficients between sNfL and combinations of diffusion-MRI metrics were even higher than the ones obtained between sNfL and the lesion load, which is known to be highly related to sNfL

levels (Chitnis et al., 2018; Todea et al., 2020). The patient, who had an extremely high level of sNfL, had a relapse 2 months before the sNfL acquisition, which may have well influenced the strong increased in sNfL levels.

To perform the correlation analyses with EDSS and sNFL, we have used subject-wise normalized maps of diffusion-based microstructural measures, which were the ones encoded by GAMER-MRI. We also trained the neural network on the original images, which, however, led to worse classification performance. Because subject-wise normalized maps were used, it is challenging to determine whether the network could learn the right pattern and to generate representative attention weights. Owing to the applied normalization procedure, the interpretation of the pathological meaning of the combined metrics is particularly difficult. Another limitation of this study was that we divided the cross-validation folds based on the number of patients instead of the number of patches: this led to different distributions of lesion and perilesional patches in the validation datasets of all cross-validation folds and to the fluctuation of the validation results. On the other hand, this also had the advantage of preventing the leak of information induced by the appearance of patches from one patient in both the training and validation dataset. Based on the obtained result (**Table 2**), the performance on the test dataset was stable, so the limitation was alleviated.

CONCLUSION

In summary, our work showed that the proposed attention-based neural network and the selection process based on the previous work can select important diffusion measures despite that they are highly intercorrelated. Those measures have the potential to be combined to enhance the correlation with the clinical measures. Future work will be required to directly find the best combinations without using a statistical test and tackling the multiple comparison problem. Furthermore, the use of a combination of diffusion-based microstructural measures deserves further attention and development, allowing a better interpretability of its pathological meaning.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the local Ethics Committee of Basel University Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

P-JL: conceptualization, data curation, methodology, investigation, formal analysis, and writing—original draft. MB: data curation, methodology, and writing—reviewing and editing. MW: resources, data curation, and writing—reviewing and editing. RR, RG, and FL: data curation and writing—reviewing and editing. SS, MBC, and AD: resources and writing—reviewing and editing. RS: conceptualization and writing—reviewing and editing. JK and LK: writing—reviewing and editing. PC: supervision and writing—reviewing and editing. CG: supervision, conceptualization, funding acquisition, resources, and writing—reviewing and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by Swiss National Funds PZ00P3_154508, PZ00P3_131914, and PP00P3_176984. FL is supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie project TRABIT (agreement no. 765148).

ACKNOWLEDGMENTS

We would like to acknowledge all the patients and healthy controls in this project. We acknowledge access to the expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Ecole polytechnique fédérale de Lausanne (EPFL), University of Geneva (UNIGE), and Geneva University Hospitals (HUG).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.647535/full#supplementary-material

REFERENCES

Akçakaya, M., Moeller, S., Weingärtner, S., and Uğurbil, K. (2019). Scan-specific robust artificial-neural-networks for k-space interpolation (RAKI) reconstruction: database-free deep learning for fast imaging. *Magn. Reson. Med.* 81, 439–453. doi: 10.1002/mrm.27420

Andermatt, S., Pezold, S., and Cattin, P. C. (2018). "Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017. Lecture Notes in Computer Science*, Vol. 10670, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Cham: Springer), 31–42. doi: 10.1007/978-3-319-75238-9_3

Andersson, J. L. R., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20, 870–888. doi: 10.1016/S1053-8119(03)00336-7

Andersson, J. L. R., and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* 125, 1063–1078. doi: 10.1016/j.neuroimage.2015.10.019

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, San Diego, CA.

Barkhof, F. (2002). The clinico-radiological paradox in multiple sclerosis revisited. *Curr. Opin. Neurol.* 15, 239–245. doi: 10.1097/00019052-200206000-00003

Barro, C., Benkert, P., Disanto, G., Tsagkas, C., Amann, M., Naegelin, Y., et al. (2018). Serum neurofilament as a predictor of disease worsening and brain and spinal cord atrophy in multiple sclerosis. *Brain* 141, 2382–2391. doi: 10.1093/brain/awy154

Behrens, T. E. J., Woolrich, M. W., Jenkinson, M., Johansen-Berg, H., Nunes, R. G., Clare, S., et al. (2003). Characterization and propagation of uncertainty in diffusion-weighted mr imaging. *Magn. Reson. Med.* 50, 1077–1088. doi: 10.1002/mrm.10609

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Cabeen, R. P., Sepehrband, F., and Toga, A. W. (2019). "Rapid and accurate NODDI parameter estimation with the spherical mean technique," in *ISMRM 27th Annual Meeting and Exhibition 11–16 May 2019*, Montreal, QC.

Chitnis, T., Gonzalez, C., Healy, B. C., Saxena, S., Rosso, M., Barro, C., et al. (2018). Neurofilament light chain serum levels correlate with 10-year MRI outcomes in multiple sclerosis. *Ann. Clin. Transl. Neurol.* 5, 1478–1491. doi: 10.1002/acn3.638

Cordero-Grande, L., Christiaens, D., Hutter, J., Price, A. N., and Hajnal, J. V. (2019). Complex diffusion-weighted image estimation via matrix recovery under general noise models. *Neuroimage* 200, 391–404. doi: 10.1016/j.neuroimage.2019.06.039

Fischl, B., Liu, A., and Dale, A. M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans. Med. Imaging* 20, 70–80. doi: 10.1109/42.906426

Fujimoto, K., Polimeni, J. R., van der Kouwe, A. J. W., Reuter, M., Kober, T., Benner, T., et al. (2014). Quantitative comparison of cortical surface reconstructions from MP2RAGE and multi-echo MPRAGE data at 3 and 7T. *Neuroimage* 90, 60–73. doi: 10.1016/j.neuroimage.2013.12.012

Granziera, C., Wuerfel, J., Barkhof, F., Calabrese, M., De Stefano, N., Enzinger, C., et al. (2020). Quantitative magnetic resonance imaging towards clinical application in multiple sclerosis. *Brain* (in press).

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE Computer Social Conference Computer Vision Pattern Recognit*, Salt Lake City, UT, 7132–7141. doi: 10.1109/CVPR.2018.00745

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. *arXiv* [Preprint]. Available online at: http://arxiv.org/abs/1802.04712 (accessed June 30, 2019).

Jelescu, I. O., Veraart, J., Adisetiyo, V., Milla, S. S., Novikov, D. S., and Fieremans, E. (2015). One diffusion acquisition and different white matter models: how does microstructure change in human early development based on WMTI and NODDI? *Neuroimage* 107, 242–256. doi: 10.1016/j.neuroimage.2014.12.009

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). Review FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Kaden, E., Kelm, N. D., Carson, R. P., Does, M. D., and Alexander, D. C. (2016). Multi-compartment microscopic diffusion imaging. *Neuroimage* 139, 346–359. doi: 10.1016/j.neuroimage.2016.06.002

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P. J., Galbusera, R., et al. (2020). Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *NeuroImage Clin.* 27:102335. doi: 10.1016/j.nicl.2020.102335

Lakhani, D. A., Schilling, K. G., Xu, J., and Bagnato, F. (2020). Advanced multicompartment diffusion MRI models and their application in multiple sclerosis. *Am. J. Neuroradiol.* 41, 751–757. doi: 10.3174/AJNR.A6484

Loshchilov, I., and Hutter, F. (2019). "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019*, La Jolla, CA.

Lu, P.-J., Yoo, Y., Rahmanzadeh, R., Galbusera, R., Weigel, M., Ceccaldi, P., et al. (2021). GAMER MRI: Gated-attention mechanism ranking of multi-contrast MRI in brain pathology. *NeuroImage Clin.* 29:102522. doi: 10.1016/j.nicl.2020.102522

Maicas, G., Carneiro, G., Bradley, A. P., Nascimento, J. C., and Reid, I. (2017). "Deep reinforcement learning for active breast lesion detection from DCE-MRI," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. Collins, and S. Duchesne (Cham: Springer), doi: 10.1007/978-3-319-66179-7_76

Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P. F., and Gruetter, R. (2010). MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *Neuroimage* 49, 1271–1281. doi: 10.1016/j.neuroimage.2009.10.002

Novikov, D. S., Fieremans, E., Jespersen, S. N., and Kiselev, V. G. (2019). Quantifying brain microstructure with diffusion MRI: theory and parameter estimation. *NMR Biomed.* 32:e3998. doi: 10.1002/nbm.3998

Reisert, M., Kellner, E., Dhital, B., Hennig, J., and Kiselev, V. G. (2017). Disentangling micro from mesostructure by diffusion MRI: a bayesian approach. *Neuroimage* 147, 964–975. doi: 10.1016/j.neuroimage.2016.09.058

Rensonnet, G., Scherrer, B., Girard, G., Jankovski, A., Warfield, S. K., Macq, B., et al. (2019). Towards microstructure fingerprinting: estimation of tissue properties from a dictionary of monte carlo diffusion MRI simulations. *Neuroimage* 184, 964–980. doi: 10.1016/j.neuroimage.2018.09.076

Rovira, Á, Wattjes, M. P., Tintoré, M., Tur, C., Yousry, T. A., Sormani, M. P., et al. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis - Clinical implementation in the diagnostic process. *Nat. Rev. Neurol.* 11, 471–482. doi: 10.1038/nrneurol.2015.106

Saha, S., Pagnozzi, A., Bourgeat, P., George, J. M., Bradford, D. K., Colditz, P. B., et al. (2020). Predicting motor outcome in preterm infants from very early brain diffusion MRI using a deep learning convolutional neural network (CNN) model. *Neuroimage* 215:116807. doi: 10.1016/j.neuroimage.2020.116807

Scherrer, B., Schwartzman, A., Taquet, M., Sahin, M., Prabhu, S. P., and Warfield, S. K. (2016). Characterizing brain tissue by assessment of the distribution of anisotropic microstructural environments in diffusion-compartment imaging (DIAMOND). *Magn. Reson. Med.* 76, 963–977. doi: 10.1002/mrm.25912

Schneider, T., Brownlee, W., Zhang, H., Ciccarelli, O., Miller, D. H., and Wheeler-Kingshott, C. G. (2017). Sensitivity of multi-shell NODDI to multiple sclerosis white matter changes: a pilot study. *Funct. Neurol.* 32, 97–101. doi: 10.11138/FNeur/2017.32.2.097

Siller, N., Kuhle, J., Muthuraman, M., Barro, C., Uphaus, T., Groppa, S., et al. (2019). Serum neurofilament light chain is a biomarker of acute and chronic neuronal damage in early multiple sclerosis. *Mult. Scler. J.* 25, 678–686. doi: 10.1177/1352458518765666

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* 23(Suppl. 1), S208–S219. doi: 10.1016/j.neuroimage.2004.07.051

Todea, R. A., Lu, P. J., Fartaria, M. J., Bonnier, G., Du Pasquier, R., Krueger, G., et al. (2020). Evolution of cortical and white matter lesion load in early-stage multiple sclerosis: correlation with neuroaxonal damage and clinical changes. *Front. Neurol.* 11:973. doi: 10.3389/fneur.2020.00973

Tournier, J. D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., et al. (2019). MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage* 202:116137. doi: 10.1016/j.neuroimage.2019.116137

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv* [Preprint]. Available online at: http://arxiv.org/abs/1706.03762 (accessed June 30, 2019).

Wattjes, M. P., Rovira, À, Miller, D., Yousry, T. A., Sormani, M. P., De Stefano, N., et al. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis - Establishing disease prognosis and monitoring patients. *Nat. Rev. Neurol.* 11, 597–606. doi: 10.1038/nrneurol.2015.157

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: convolutional block attention module," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Cham: Springer), 3–19. doi: 10.1007/978-3-030-01234-2_1

Ye, Z., George, A., Wu, A. T., Niu, X., Lin, J., Adusumilli, G., et al. (2020). Deep learning with diffusion basis spectrum imaging for classification of multiple sclerosis lesions. *Ann. Clin. Transl. Neurol.* 7, 695–706. doi: 10.1002/acn3.51037

Yoo, Y., Tang, L. Y. W., Brosch, T., Li, D. K. B., Kolind, S., Vavasour, I., et al. (2018). Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. *NeuroImage Clin.* 17, 169–178. doi: 10.1016/j.nicl.2017.10.015

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). "Image captioning with semantic attention," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, doi: 10.1109/CVPR.2016.503

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* 61, 1000–1016. doi: 10.1016/j.neuroimage.2012.03.072

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

METHODS
published: 15 April 2021
doi: 10.3389/fnins.2021.633438

Management and Quality Control of Large Neuroimaging Datasets: Developments From the Barcelonaβeta Brain Research Center

Jordi Huguet[1], Carles Falcon[1], David Fusté[1], Sergi Girona[2], David Vicente[2], José Luis Molinuevo[1], Juan Domingo Gispert[1] and Grégory Operto[1]* for the ALFA Study*

[1] Barcelonabeta Brain Research Center, Barcelona, Spain, [2] Barcelona Supercomputing Center, Barcelona, Spain

OPEN ACCESS

Edited by:
Meritxell Bach Cuadra, Lausanne University Hospital (CHUV/UNIL), Switzerland

Reviewed by:
Oscar Esteban, Stanford University, United States
Karl Helmer, Massachusetts General Hospital, Harvard Medical School, United States
Jo Etzel, Washington University in St. Louis, United States

***Correspondence:**
Juan Domingo Gispert
jdgispert@barcelonabeta.org
Grégory Operto
goperto@barcelonabeta.org

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 25 November 2020
Accepted: 02 March 2021
Published: 15 April 2021

Citation:
Huguet J, Falcon C, Fusté D, Girona S, Vicente D, Molinuevo JL, Gispert JD and Operto G (2021) Management and Quality Control of Large Neuroimaging Datasets: Developments From the Barcelonaβeta Brain Research Center. Front. Neurosci. 15:633438. doi: 10.3389/fnins.2021.633438

Recent decades have witnessed an increasing number of large to very large imaging studies, prominently in the field of neurodegenerative diseases. The datasets collected during these studies form essential resources for the research aiming at new biomarkers. Collecting, hosting, managing, processing, or reviewing those datasets is typically achieved through a local neuroinformatics infrastructure. In particular for organizations with their own imaging equipment, setting up such a system is still a hard task, and relying on cloud-based solutions, albeit promising, is not always possible. This paper proposes a practical model guided by core principles including user involvement, lightweight footprint, modularity, reusability, and facilitated data sharing. This model is based on the experience from an 8-year-old research center managing cohort research programs on Alzheimer's disease. Such a model gave rise to an ecosystem of tools aiming at improved quality control through seamless automatic processes combined with a variety of code libraries, command line tools, graphical user interfaces, and instant messaging applets. The present ecosystem was shaped around XNAT and is composed of independently reusable modules that are freely available on GitLab/GitHub. This paradigm is scalable to the general community of researchers working with large neuroimaging datasets.

Keywords: processing workflows, neuroimaging, quality control, data management, neuroinformatics, cohort studies

INTRODUCTION

Neuroimaging has now taken a central role in the context of research in Alzheimer's disease (AD) as in neuroscience in general. Its non-invasive nature, its relative widespread availability, and its potential to provide efficient disease predictive markers have incentivized global efforts to assemble large imaging datasets, with numbers of subjects starting to reach ranges of epidemiological studies (Van Horn and Toga, 2014; Abe et al., 2015; Júlvez et al., 2016; Miller et al., 2016; Cox et al., 2019). With the advent of modern computational methods and the constant progress in imaging techniques, images are now routinely taken through automatic processing workflows, yielding a

series of endpoints to be analyzed against other variables, which may potentially develop into findings. Despite good practices and quality assurance (QA), each step (acquisition or processing) is likely to exhibit anomalous behaviors and may lead to erroneous conclusions if unnoticed. In this regard, quality control (QC) protocols are designed to track down and protect against such errors but have until now faced major obstacles. Their purpose is to assess the conformity of any applicable dataset with a set of custom specifications and consequently determine whether the dataset is suited for further processing/analysis. On the one hand, individual visual inspection has proven to be neither fail-safe nor compatible with the size of the largest cohort studies (Alfaro-Almagro et al., 2018). On the other hand, automated or semi-automated QC offers promising cost-reducing perspectives (Esteban et al., 2019a; Sunderland et al., 2019); however, it remains hard to generalize as it strongly depends on the study design (single/multisite, clinical/cohort study) and needs to be adapted to each imaging sequence (Oguz et al., 2014; Bastiani et al., 2019) and each step of the workflow (raw images, processing outputs) (Klapwijk et al., 2019). **Table 1** draws an inventory of existing resources focused on QC of neuroimaging data, automated or not, with corresponding references and repositories, if applicable. This list is first and foremost illustrative of their variety and specificity in relation to types of input data. Interestingly, the recent years have seen the emergence of new approaches aiming at unifying, on one side, QC protocols across groups and, on the other, processing workflows in some of these modalities such as structural magnetic resonance imaging (MRI) (Esteban et al., 2017) or functional MRI (Esteban et al., 2019b). Such approaches may pave the way for a general process of standardization of QC tools and procedures that would extend to most used neuroimaging data modalities.

l Improved data management is also directly associated with improved quality assessment: a system in which one can easily find and work with the data is likely to make quality assessment easier. Inversely, a system in which finding the data is complicated will make quality assessment much harder. As a consequence, the capacity to evaluate the results of any workflow and the capacity to identify/navigate through them in a larger repository are both tightly coupled. This is especially relevant for workflows such as the ones used in neuroimaging studies, which typically combine high levels of complexity, heterogeneity (e.g., in numbers of files, nature/structure of data) on the one hand, and, on the other, a high degree of required expertise to assess their outputs. With respect to this, to date, individual research groups may choose among different strategies, essentially based on their size and allocated resources, among which:

- organizing a local file repository and relying on core tools/libraries, predefined procedures and adoption of best practices.
- setting up a local management platform by building upon some existing open-source or proprietary systems (or developing it from scratch).
- subcontracting data management as a service, as included in "Science in the cloud" solutions.

Different sets of technical solutions exist for each of these approaches. In particular, initiatives such as BIDS (Gorgolewski et al., 2016) or BIDS-Apps (Gorgolewski et al., 2017) play an extremely valuable role in the spread of software-engineering practices along the neuroimaging research workflow, with beneficial consequences on reproducibility. The BIDS standard has become, over the past years, a spearhead in the promotion of FAIR principles (Wilkinson et al., 2016) by addressing data *findability*, *reusability*, and *interoperability* across groups, systems, and tools. As BIDS provides the formalism to organize the data and metadata, data *accessibility*, for its part, requires additional software that will generally include basic features for data management and exploration. As two open-source cloud-based solutions that have built upon BIDS, OpenNeuro (Poldrack et al., 2013) and Brainlife.io (Avesani et al., 2019) are iconic examples of platforms giving access not only to datasets but also to online computational resources, giving substance to the concept of virtual laboratory (Frisoni et al., 2011). As such, the purpose of the "Science in the cloud" model is also to facilitate data sharing and reproducibility by centralizing resources for data storage, management, computation, and QC in the neuroinformatics field. This model has begun to spread (Redolfi et al., 2015; Manjón and Coupé, 2016; Kiar et al., 2017; Glatard et al., 2018) and draws a promising future for the community. Notwithstanding the preceding, it may still fail to address immediate down-to-earth needs from small to average-sized research groups, especially the ones dealing with self-acquired imaging data. First, implementing these frameworks or adapting them locally requires strong IT skills and a specialized labor force, making it technically out of reach for many groups with insufficient human and/or computational resources, or without connection to large consortia. Second, relying on existing open-access instances is still hardly compatible with data confidentiality policies in most studies, as these are rarely permissive enough to allow upload to third-party platforms from the start. The basic needs of the many research groups include, for instance, basic data collection/querying/handling in average-sized datasets (e.g., up to several thousands of subjects), combined with further exploration/review along most typical analysis workflows. It is particularly compelling that in comparison to the magnitude of efforts underway to assemble large imaging datasets, the range of technical solutions to address such basic needs is actually limited. As previously reported by Nichols and Pohl (2015) and Shenkin et al. (2017), extensible neuroimaging archive toolkit (XNAT) (Marcus et al., 2007), LORIS (Das et al., 2010), and NIDB (Book et al., 2013) appear indeed as the main existing open-source neuroinformatics software platforms supporting data sharing.

Now that neuroscience has entered a propitious era of data and computation, practical solutions are still required to efficiently operate local databases and run tailored controls on complex type-agnostic raw and processed data.

Quality control and data management are thus both interrelated. They both have transversal impacts on the research workflow, from the data acquisition to the analysis. Both if poorly executed may have a strong negative impact on reproducibility.

TABLE 1 | List of currently available resources intended for quality control of neuroimaging data (adapted from https://incf.github.io/niQC/tools).

Name	References	Data	Technology	Code repository
dashQC	n/a	fMRI, registration	Javascript	https://github.com/SIMEXP/dashQC_fmri/issues
qcApp	n/a	FreeSurfer	Java	https://github.com/ntraut/QCApp
qsiprep	n/a	DWI	Python	https://github.com/pennbbl/qsiprep
uniQC	n/a	fMRI	Matlab	https://github.com/CAIsr/uniQC
exploreDTI	Leemans et al., 2009	DWI	Matlab	n/a
dtiprep	Oguz et al., 2014	DWI	C++	https://github.com/NIRALUser/DTIPrep
PCP-QAP	Shehzad et al., 2015	T1w, fMRI	Python	https://github.com/preprocessed-connectomes-project/quality-assessment-protocol
brainbox	Heuer et al., 2016	segmentation	Javascript	https://github.com/OpenNeuroLab/BrainBox
exploreASL	Mutsaerts et al., 2017	ASL	Matlab	n/a
mriqc	Esteban et al., 2017, Esteban et al., 2019a	T1w, fMRI	Python	https://github.com/poldracklab/mriqc
PALS	Ito et al., 2018	T1w, fMRI	Python	https://github.com/npnl/pals
rtQC	Heunis et al., 2019	fMRI	Matlab	https://github.com/rtQC-group/rtQC
visualqc	Raamana, 2018	T1w, FreeSurfer	Python	https://github.com/raamana/visualqc
mindcontrol	Keshavan et al., 2018	FreeSurfer	Python, Javascript	https://github.com/OpenNeuroLab/mindcontrol
AFQ-Browser	Yeatman et al., 2018	DWI	Python, Javascript	https://github.com/yeatmanlab/AFQ-Browser
braindr (braindrles)	Keshavan et al., 2019	snapshots	Javascript	https://github.com/OpenNeuroLab/braindr; https://github.com/SwipesForScience/SwipesForScience
eddyqc/quad/squad	Bastiani et al., 2019	DWI	C (FSL)	https://git.fmrib.ox.ac.uk/matteob/eddy_qc_release
fmriprep	Esteban et al., 2019b	fMRI	Python	https://github.com/poldracklab/fmriprep
qoala-t	Klapwijk et al., 2019	FreeSurfer	R	https://github.com/Qoala-T/QC
snaprate	Operto, 2019	snapshots	Python, Javascript	https://github.com/xgrg/snaprate
nisnap	Operto and Huguet, 2020	snapshots	Python	https://github.com/xgrg/nisnap

As advocated in the neuroimaging community, e.g., by the ReproNim initiative (Kennedy et al., 2019), core resources may already exist but their use should be facilitated so as reproducibility is achieved by design, not as an afterthought. Such considerations have nurtured the development of a novel infrastructure scheme–presented here–for imaging data management and processing, focused on facilitating scalable QC and aiming at maximizing the reuse of existing open core tools/libraries.

This model was implemented and adapted to the needs of a specific research program, namely, the ALFA project, yet with concerns about lean development principles and reusability. The ALFA project (Alzheimer's and Families) is a research platform started by the Barcelonaβeta Brain Research Center (BBRC) for the prospective follow-up of a cohort of cognitively normal subjects–most of which are the offspring of AD patients. Extensive phenotyping of participants includes cognitive assessment, lifestyle questionnaires, blood extraction for further genetic analysis, cerebrospinal fluid collection, positron emission tomography (PET) imaging, and multimodal MRI examination performed on-site on a single Philips Ingenia CX 3T scanner. The interested reader may refer to Molinuevo et al. (2016) for a full description of the various arms of the project and administered examinations. Since 2012 when BBRC was created, its neuroimaging platform has been acquiring and is currently managing data from over 5000 participants across its different studies. Imaging protocols include standard MRI sequences (with T1/T2/diffusion-weighted, inversion recovery, and resting-state functional MRI), some more advanced ones (arterial spin labeling, susceptibility-weighted imaging, and quantitative flow, among others), and, for a subset of participants, PET imaging–fluorodeoxyglucose (FDG) and flutemetamol. This paper documents the core concepts and implementation of this infrastructure for imaging data management, processing, and control. The first section will detail the routine data flow at BBRC, which this infrastructure partially supports. In a second section, the paper will describe the different ways provided to researchers of the group to interact with the platform. The third section will focus on QC performed on large imaging datasets. The fourth section will then elaborate on the employed strategy to foster sustainability and reproducibility and describe principles for future development.

BBRC: ANATOMY OF A SINGLE-SITE IMAGING RESEARCH PLATFORM

Participants may be included in one of the hosted programs such as the ALFA study, and get assigned with a unique accession number (**Figure 1**). This accession number is represented as a barcode and follows the participant through the whole acquisition protocol, which, on a standard basis, includes full neuropsychological evaluation, assessment of clinical history, APOE genotyping, lifestyle questionnaires, blood sampling, and–for a subset of individuals–cerebrospinal fluid extraction. Structural and functional MRI is acquired on-site on a dedicated MR scanner. Participants of the ALFA+ program undergo both flutemetamol and FDG PET at the Hospital Clinic of Barcelona.

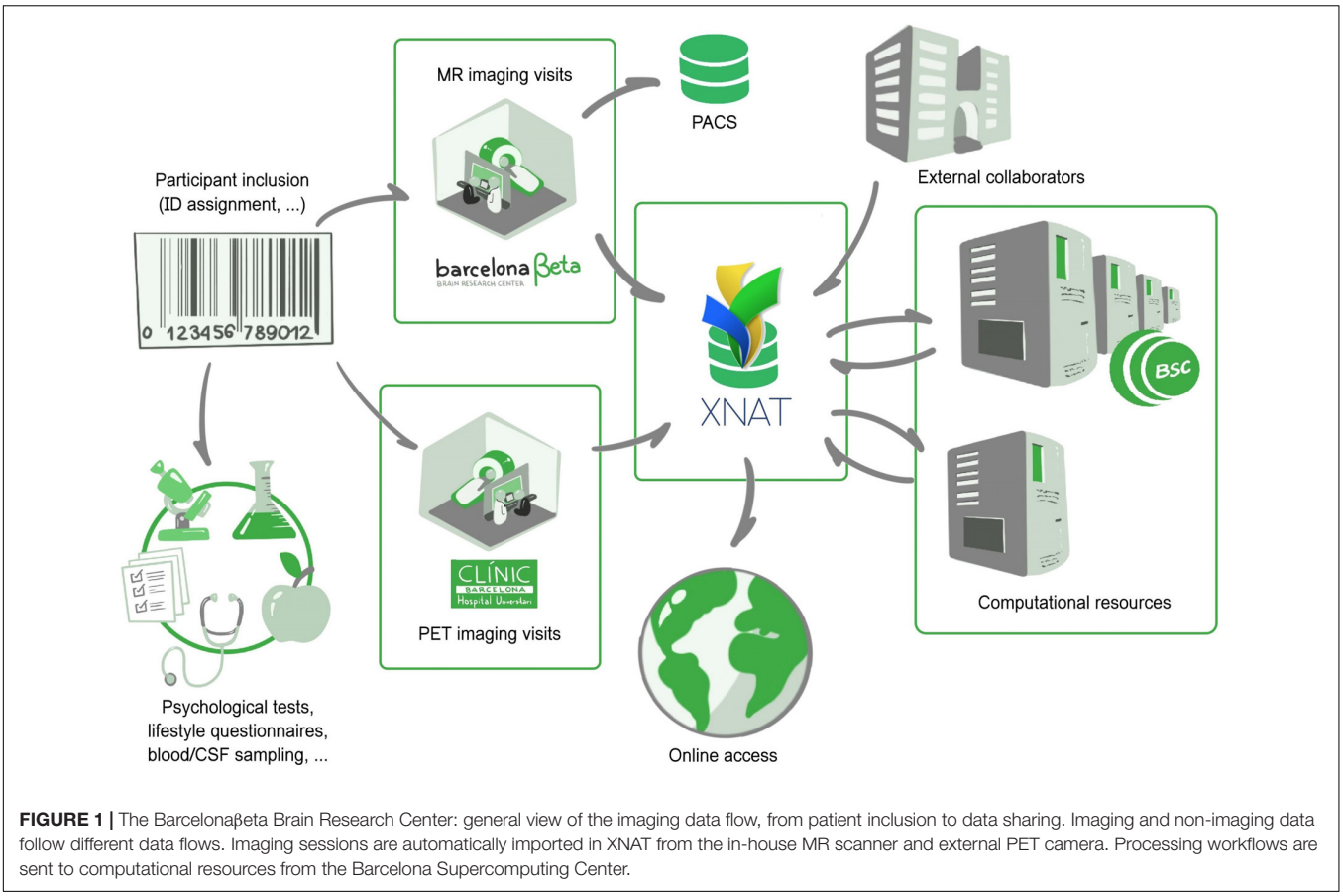


FIGURE 1 | The Barcelonaβeta Brain Research Center: general view of the imaging data flow, from patient inclusion to data sharing. Imaging and non-imaging data follow different data flows. Imaging sessions are automatically imported in XNAT from the in-house MR scanner and external PET camera. Processing workflows are sent to computational resources from the Barcelona Supercomputing Center.

Imaging and non-imaging data are stored and managed in two individual platforms. Non-imaging data are imported into a relational database and follow a specific data flow that is not described here. Imaging data are directly transferred from the scanner to both a PACS archive and an XNAT platform. Extensible neuroimaging archive toolkit (XNAT) (Marcus et al., 2007) is the most broadly deployed open source system to have emerged among imaging platforms in recent history. In this context, the PACS archive is used for long-term backup purposes, preserving a pristine copy of the acquired imaging data, and for daily routine visual review and reporting by radiologists, whereas XNAT is a much more flexible system geared toward researchers, allowing transformation, automatic processing, browsing, downloading, and eventually sharing. A Clinical Trial Processor (CTP) service (Aryanto et al., 2012) is run between the MR scanner and XNAT to ensure proper de-identification of protected health information. Outsourced PET imaging data are directly pulled from the acquisition site: a daily daemon service pulls new imaging scans from an sFTP server and pushes them to the PACS archive which then auto-forward to XNAT (via CTP). The workflow is open to external collaborators, who may also push data in independently managed projects distinct from the ALFA study.

Once the data have been successfully imported into XNAT, imaging sessions are routed to their corresponding XNAT project/study and then taken through automatic workflows. These workflows are managed by the XNAT Pipeline Engine, which directly draws computational power from the Barcelona Supercomputing Center[1]. Workflows include processing–e.g., involving all types of neuroimaging software or published methods/algorithms–but also automatic controls based on *Validators*, as described further in section "Generalized Automatic Sanity Check/Quality Control." This results in the generation of derived images, numerical endpoints, or validation reports. Along with the primary raw data, they form the body of online available resources that users may reach by then logging into the system.

This data flow is presented in **Figure 1**.

BETTER CONTROL ON DATA BY PROVIDING MULTIPLE ACCESS WAYS

XNAT as the Infrastructure Core Engine for Imaging Data Management

Among the most significant ones from the last decade, neuroimaging projects like the Open Access Series of Imaging Studies (Marcus et al., 2010), IMAGEN (Schumann et al., 2010), the Human Connectome Project (Marcus et al., 2011), the International Neuroimaging Data-sharing Initiative (Mennes et al., 2013; Kennedy et al., 2016), the Adolescent Brain Cognitive Development (Casey et al., 2018), the UK Biobank (Miller et al.,

[1]http://www.bsc.es

2016), followed by the more recent ONDRI (Scott et al., 2020) or EPAD (Ritchie et al., 2020), have all in common that their respective infrastructures for data sharing are based on XNAT. This not only confirms the status of XNAT as a central technology but also highlights the opportunity of any model built around XNAT in terms of reusability.

We chose to rely on XNAT as the core engine of our infrastructure for imaging data. Among the few existing options available, XNAT offers an adequate cost–benefit ratio for groups of all sizes when comparing the complexity of implementation to all of its built-in features. XNAT provides tools for common management, user access, data processing, and sharing, thus covering many aspects of the basic neuroimaging workflow. It also includes a DICOM storage service (C-STORE SCP) for receiving and sorting images from any DICOM-compliant imaging device, which is essential for organizations managing their own imaging equipment. User access to the archive is provided by a secure web application. Workflow execution is enabled by a Pipeline Engine, while XNAT maintains full histories by tracking all changes to the data, thus enforcing data traceability. Finally, XNAT implements a security system that allows administrators to grant access to specific actions or datasets following predefined user roles.

To date, XNAT is still under active development with strong community-based support, aligning with current trends in the community as shown by recent support for BIDS format and containerized data processing (e.g., using Merkel, 2014). Most users may operate the database and search the repository through the built-in web-based application. Aside from this graphical interface, XNAT provides a Representational State Transfer (REST) Application-Program Interface (API) that allows users to query the database and therefore programmatic interaction with its contents. Furthermore, the *pyxnat* (Schwartz et al., 2012) library capitalizes on this API and allows users to interact with XNAT using Python.

We advocate that users should have multiple proposed ways and be free to choose their preferred one to operate the platform, as a greater flexibility in this regard is a stepping stone for improved data review and issue tracking. With respect to this, a few previous examples have built onto XNAT (Gee et al., 2010; Harrigan et al., 2016; Job et al., 2017), often leveraging its RESTful API (Schwartz et al., 2012; Gutman et al., 2014), to extend its standard features and present new ones. Such an approach stands out by its light footprint, relying on XNAT's core features without needing to touch its codebase, to the mutual benefits of maintainability, dependability, portability, and usability. In line with this approach, this present paper describes a collection of lightweight solutions which together form an adaptive modular ecosystem focused on user experience and neuroimaging data QC.

Barcelonaβeta + XNAT: *bx*

Interacting with the data on XNAT can be done mainly in two ways: either graphically using the web application or through a REST API. While the former is suited for all profiles, the latter is intended for a more technical category of users, allowing them to automate bulk operations, e.g., downloading large collections of data and populating projects or any type of systematic task that would otherwise, using the web application, require many manual operations. Version 1.7.5 of XNAT now includes a Desktop Client that may be used to download collections of images for instance from an entire study (or *project* in XNAT jargon). Still, between "all clicks" and "all script" lies a large gray zone with users who without being experienced coders may still have some knowledge on how to use command-line tools. For this special category, we wrote *bx*, which allows us to run from a terminal among a predefined set of bulk operations using a single command. This includes, for instance:

- downloading images of a given sequence over a project in the NIfTI format (better suited to a majority of post-processing software suites).
- downloading processing outputs over a project (e.g., segmentation maps, 3D models, etc.).
- downloading an Excel table with all numeric outcomes from a given pipeline over a project.
- downloading a table with acquisition dates from an entire project.
- in general, downloading any given type of resources over an entire project.

In particular, to get a local copy of the results from FreeSurfer *recon-all* pipeline (Fischl, 2012) over the entire XNAT project *ALFA*, one would simply run:

```
bx freesurfer6 files ALFA
```

Destination folder is set in a locally stored configuration file along with the user's XNAT login credentials.

By extension, the following command:

```
bx freesurfer6 aseg ALFA
```

would generate a single spreadsheet file containing all the structural volumes estimated by FreeSurfer (in aseg.stats files). The current version (0.1.6) also includes, among others, commands for SPM (Ashburner and Friston, 2005), ANTs (Avants et al., 2009), FSL (Jenkinson et al., 2012), ASHS (Yushkevich et al., 2015), and CAT (Gaser, 2016), with subcommands for collecting output files, measurements yielded by the pipeline, QC-oriented snapshots, validation reports, or automatic test outcomes (as described later in section "Generalized Automatic Sanity Check/Quality Control"). Importantly, any command may be applied to an entire project, one single MRI session, or also curated image collections[2] relying on discretionary criteria (e.g., based on clinical, genetic or cognitive characterization, or any other external variable).

Such a tool thus provides an additional command-based way to interact with the XNAT data which optimizes a set of "frequent" use cases (based on user reports, like bulk downloading pipeline outputs) while abstracting the rest (i.e., obviating intermediate steps such as selection of subjects/experiments/resources). Since it was built over *pyxnat*, this makes it rather easy to get adapted to specific local configurations (or additional resources).

[2]As inventoried by the command: `bx lists`.

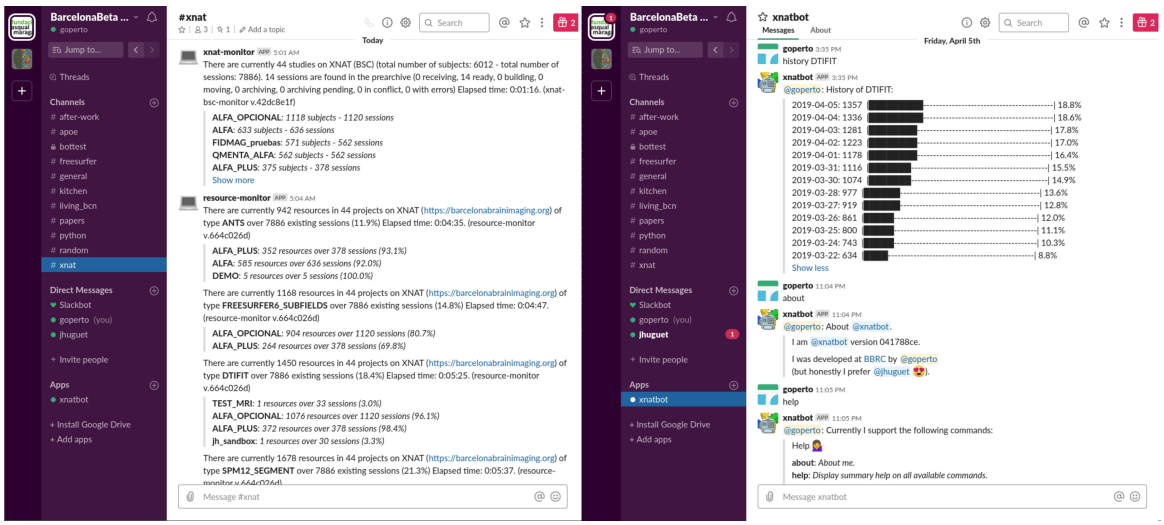


FIGURE 2 | Screenshots of the #xnat channel from the Barcelonaβeta Slack workspace. **(Left)** Monitors provide members of the channel with daily updates on the current data available on the imaging platform without any user action. **(Right)** Basic human chatbot interactions give access to more specific statistics. In this example, the user is querying for the progress over time of some processing task (with DTIFIT).

It is distributed as a PyPI package under the name *bbrc-bx* and hosted on GitLab: https://gitlab.com/xgrg/bx.

Cron Jobs, Bots, and Monitors

In addition to *bx*-like scripts and XNAT's standard interface, daily summaries are delivered automatically through both emails and instant messaging (IM). We built onto XNAT email notification service so that subscribed users receive a comprehensive sanity report (detailed in section "Generalized Automatic Sanity Check/Quality Control") for every new session uploaded from the scanner. In parallel, automatic monitors running on a Slack (Johnson, 2018) #xnat channel provides authorized users with daily updates on numbers of available subjects/raw sessions per project and available resources such as processing outputs (derivatives) (left part in **Figure 2**). Such automatic delivery systems complement standard user experience by directly feeding with periodic statistics on the database, thus allowing to check instantly on the system's general integrity status without user action. Users may also get further customized views on this information through basic human-chatbot interactions, e.g., longitudinal statistics. **Figure 2** illustrates this integration: on the left, members of the #xnat channel are updated every day on available data, and on the right, users may ask about the progress over time (daily numbers of a given resource) of any pipeline on the platform. This approach may naturally be adapted to other messaging systems (e.g., Mattermost, Riot, Zulip, IRC) or project management tools possessing an API (e.g., Trello, Basecamp).

We advocate for giving users multiple controlled ways to deal with data. XNAT RESTful API is one of the most powerful features of its framework and allows to build a variety of access modalities, each of which comes with pros and cons. For example, the graphical user interface gives individual and comprehensive control on the data, though manually operated; *pyxnat* adds a programmatic interface to it and is, therefore, rather developer-oriented; *bx* optimizes bulk downloading operations from scripts, yet for a set of pre-selected resources; and IM-based tools provide only high-level summarized information but add an interactive and collaborative touch and nicely intertwine with natural conversations among users.

"GIVEN ENOUGH EYEBALLS, ALL GLITCHES ARE SHALLOW"[3]

Each step of an analysis workflow should ideally be paired with specific checkpoints. Given the increasing quantity and complexity of datasets, relying on automatic control is imperative, but manual inspection can rarely be avoided. The following approach aims at capitalizing on automatic controls while allowing multiple users to jointly participate in visual inspection.

Generalized Automatic Sanity Check/QC

In line with recent trending standards such as BIDS (Gorgolewski et al., 2016) or NIDM (Keator et al., 2013), we present a validator-based modular approach, which, in our current implementation, covers image types such as T1-weighted, DWI, and PET images, and processing outputs like FreeSurfer, SPM, and FSL DTIFIT, and the approach may be easily extended to others. For each type of data, we define a tailored procedure for QC. Such procedures consist of predefined sequences of checkpoints: each checkpoint (later referred to as *Test*) is associated with some particular aspect of the data and would result either as passed or failed. In this present implementation, every new imaging resource pushed into the system is thus automatically taken through a QC

[3]Adapted from Raymond (1999)

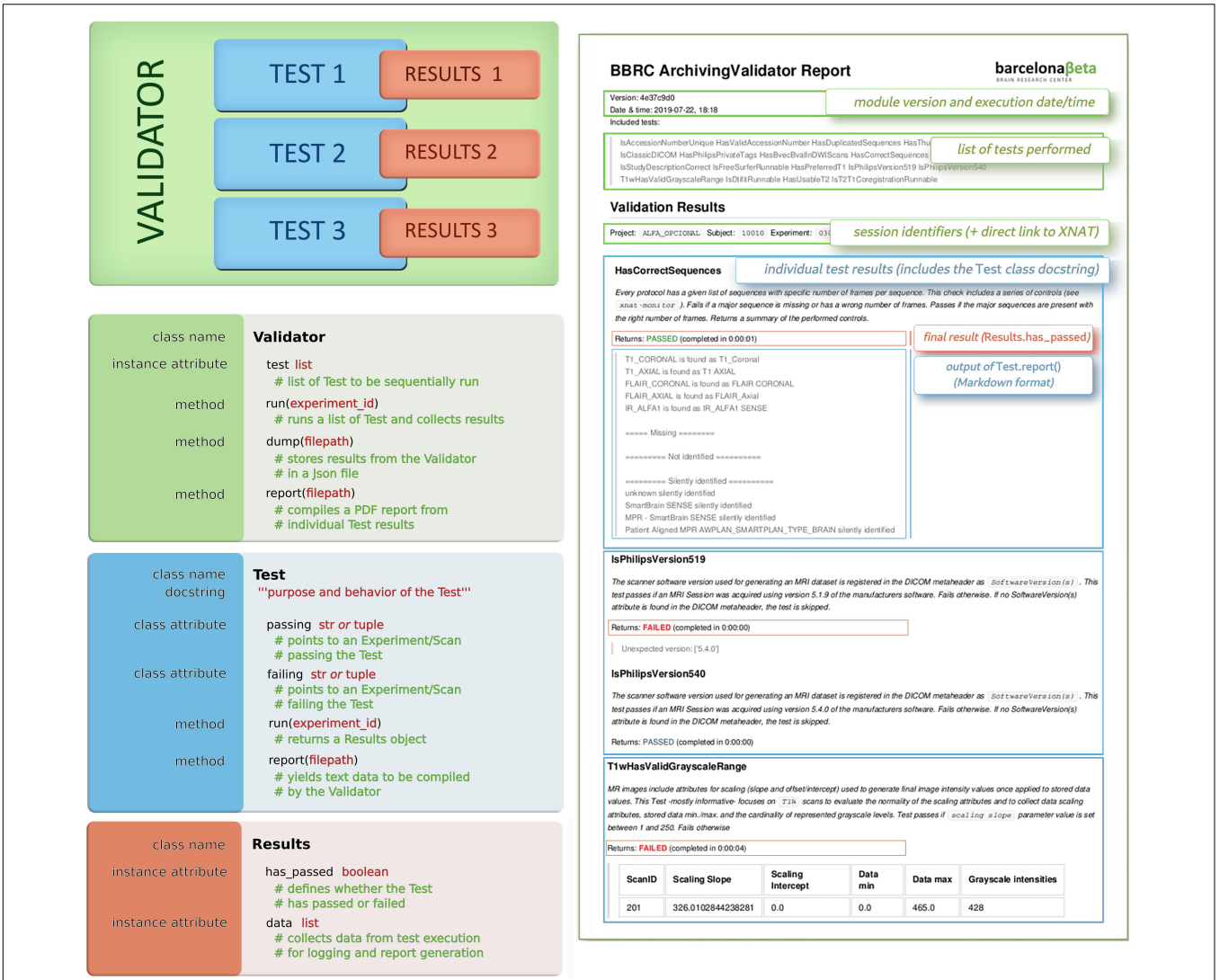


FIGURE 3 | (Left) *Validators*: concepts and classes. *Validators* and *Tests* all share the same template. *Validators* are defined by a list of *Tests*, which in turn yield some *Results*. Each *Results* object embeds a main Boolean, which defines whether the *Test* was successful, and some additional data for logging purposes or report generation. **(Right)** Example of a produced validation report (only the first page is displayed); the color code highlights the matching between sections of the report and the corresponding concepts: green refers to *Validators*, each blue area corresponds to a *Test*, and *Results* are shown in red squares.

procedure adapted to the type of data. Checkpoints are defined based on aspects of the data or metadata known to potentially exhibit undesired variability, e.g., due to technical or human-related factors. They may, for instance, include verifying that the output of some process matches some expected list of files, that some image parameters fall in specific intervals. Nevertheless, the approach is designed so as to give the most flexibility and scalability to the range of possible checkpoints. The use of a single template for all checkpoints–each of them being documented with human-readable specifications (e.g., detailed in each *Test*'s *docstring*, as explained hereafter) and resulting in a binary outcome–makes them easier to read and comprehend, especially in code. As a result, every new imaging session is provided with a checklist, by which the execution of further pipelines may be conditioned. It is worth noting though that by being designed as an independent command-line tool, any procedure from this module may be executed, not only automatically, but also manually upon request on any applicable dataset. The tool, written in Python, is based on two nested concepts: *Tests* and *Validators* (**Figure 3**).

A *Validator* is an object defined by a set of *Test* objects, each of which would check specific traits of a given XNAT entity (e.g., an incoming imaging session, or results from a processing workflow). *Validators* are run like any other pipelines by XNAT Pipeline Engine, triggered by some functional events (e.g., archiving of a session and completion of a processing pipeline, among others). The outputs from these series of checks are stored as additional resources and would be used to infer, either by visual review or programmatically, on the validity of the target resource.

A *Test* is defined by a *run()* and a *report()* function. The *run()* function returns a *Results()* object that has two attributes, namely, *has_passed* (Boolean) and *data* (list). This *run()* function may target any resource, either an Experiment or a Scan (following the XNAT terminology). Every *Test* has also two hardcoded class-level attributes, namely, *passing* and *failing*, pointing at two Experiments (or two 2-uples Experiment + Scan) from the running XNAT instance on which the test should respectively pass and fail [used for continuous integration (CI)]. Depending on the test purpose, it may return *Results*(*has_passed* = True) or *Results*(*has_passed* = False). One additional *data* argument may be passed to the *Results* constructor to record extra information (e.g., elapsed time) from the test execution.

In practice, running a *Validator* on a given experiment takes its associated set of *Tests* and runs them sequentially. A *Test* may apply to a Scan instead of an Experiment (e.g., *checking that DICOM files have been converted to NIfTI*), in which case the *Test* could be performed over all the existing Scans of the Experiment. Upon failure of a *Test*, scan quality flags may be adjusted from usable to questionable/not usable on XNAT. Once completed, the *Validator* dumps the results data in a JSON record and generates a Markdown-based PDF report (**Figure 3**). This report is built by calling each *Test*'s *report()* function consecutively and compiling their results in as many individual sections. By default, every section includes the *docstring* attribute taken from every *Test* class for the sake of traceability and self-sufficiency.

Both resulting PDF and JSON files appear on XNAT as resources of the validated experiment, so that users may query on them[4] or dump them from the entire database, e.g., into a single spreadsheet file[5]. This is made directly possible using *bx* commands (section "Barcelonaβeta + XNAT: *bx*") thanks to the seamless integration between both tools.

One key strength of this model is its adaptability/genericity. It allows rapid implementation of new *Tests* on any type of imaging data provided it can be identified as an XNAT Experiment or Scan. The actual performed verifications are stated in the *run()* function and may hence use any required external library. Another key advantage is the low cost associated with CI-related maintenance. Regression testing is indeed critical for the system to be sustainable as more checkpoints and more data are added. Automated unit testing for CI is performed after every new change in the code, based on the two class attributes *passing* and *failing* provided for each *Test*. Every single *Test* is thus systematically re-executed against two specific cases after any change in the code. Along with this, each generated report includes a reference to the last SHA identifier issued by the version control system. As all Tests share the same template, the testing code for CI requires no updates and remains always adapted to any newly added *Test*. Such a design yields to a unit-test-to-production-code ratio currently under 1:30.

In our current implementation, *Tests* have so far covered aspects related to both MR and PET acquisition and their post-processing derivatives. **Supplementary Table 1** gives an illustrative summary of currently implemented *Tests*, including their associated docstrings to describe their purpose.

For example, every time a new PET session is imported to XNAT, a *PetSessionValidator* is triggered. This *Validator* currently includes a set of nine Tests. The first one, *IsTracerCorrect*, checks that the tracer information is correctly registered in the DICOM headers. The second one, *IsSeriesDescriptionConsistent*, makes sure that metadata are consistent across the session; then, *IsScannerVersionCorrect* checks in the DICOM headers that the scanner model matches, in this case, "SIEMENS Biograph64 VG51C". Then, follow *IsSubjectWeightConsistent* and *IsTracerDoseConsistent* controlling that the values registered for subject's weight and tracer dose match some target intervals (between 40 and 150 kg and between 1.5e8 and 3.5e8 Bq, respectively). Finally, the *Validator* runs *IsSubjectIdCorrect* to ensure the subject's ID has the right format; *HasUsableT1*, which checks whether the subject has a valid T1-weighted image stored on XNAT; and both *IsCentiloidRunnable* and *IsFDGQuantificationRunnable*, which assess whether the data are suited for the execution of two quantification pipelines.

Another example is *ASHSValidator*, which is triggered every time some hippocampal subfield segmentation is executed over an MR session (using the ASHS pipeline). The *Validator* sequentially runs *HasAllSubfields*, which makes sure that all expected subfields appear in the final segmentation; *HasCorrectASHSVersion* controlling the software version; *HasCorrectItems* checking that the list of generated files matches the right one; *HasNormalSubfieldVolumes*, which assesses whether resulting subfield volumes fall inside some safety intervals; and *ASHSSnapshot*, which generates a snapshot of the final segmentation (shown in **Figure 4**).

Other *Validators* include, for instance, *ArchivingValidator* (triggered every time an MR session is imported/archived), *SPM12Validator*, *CAT12Validator*, *FreeSurfer6Validator*, *ANTSValidator*, and *DTIFITValidator* (triggered after every execution of SPM12, CAT12, FreeSurfer6, ANTS, and FSL DTIFIT, respectively). For a more comprehensive list of *Tests*, *Validators*, and details on their purpose, the reader may refer either to **Supplementary Table 1** or directly to the code repository for the latest version, as sharing the same template [where each *Test* is a class with two test cases, a docstring, a *run()*, and a *report()* function, as described above] makes them easily readable.

The source code is released as an independent tool, *bbrc-validator*, available as a PyPI package and code is hosted on GitLab[6].

Generating Summarized Representations of Segmentation Results: *nisnap*

Among the broad typology of outputs generated by most neuroimaging analysis workflows, numeric and image-based results are probably the most common. In particular, any segmentation technique will generally yield either a label volume

[4]For instance, using the following bx command: `bx spm12 report <project>` (to download reports from SPM12 segmentations from a whole XNAT project).

[5]The command would be: `bx spm12 tests <project>`.

[6]https://gitlab.com/bbrc/xnat/bbrc-validator

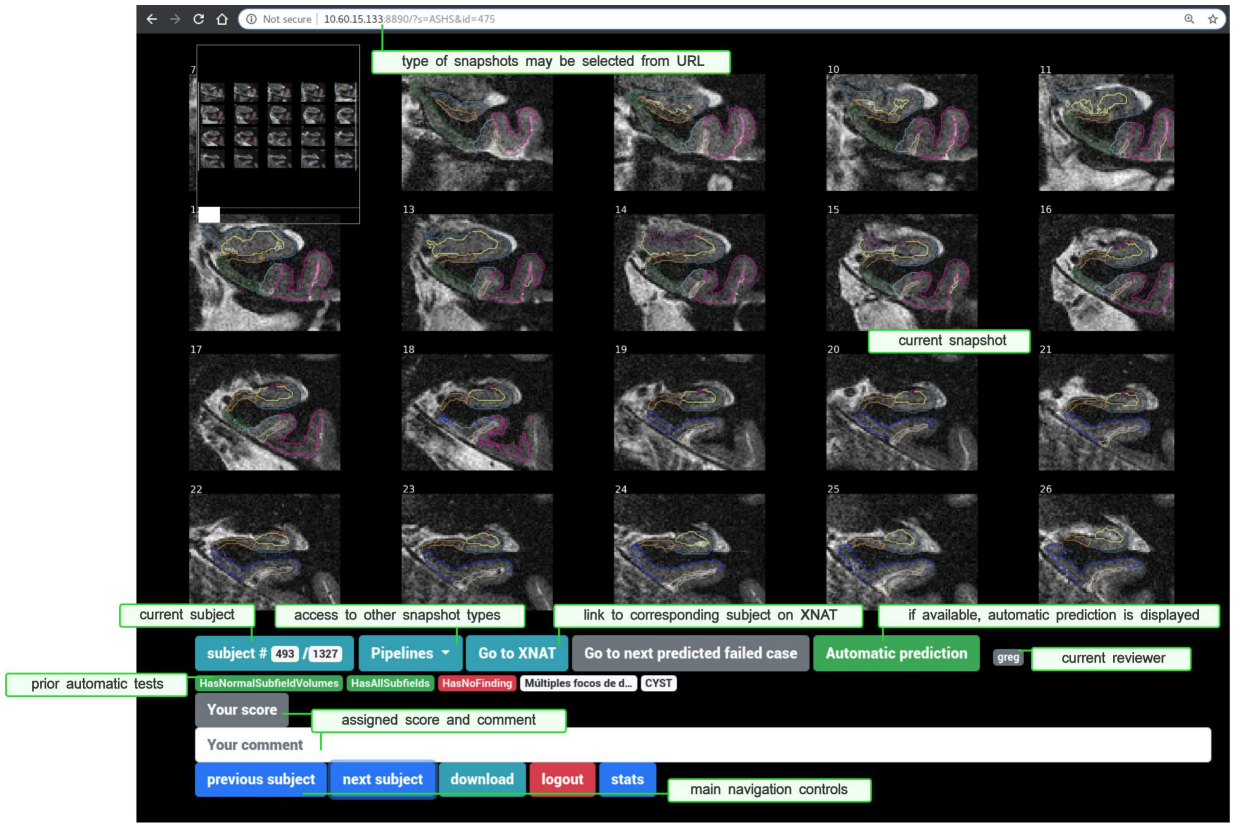


FIGURE 4 | Snaprate: General user interface, running in a web browser. The upper part displays a zoomable snapshot (here a segmentation of hippocampal subfields). The lower part shows a section for the review section and navigation controls, including links to XNAT and to other types of snapshots. Results from prior checkpoints are also displayed in red (if failed) or green (if passed).

or probability maps to describe some target structures/objects, possibly coming with some derived numeric descriptors, as this is the case with standard software such as SPM or FreeSurfer for cortical/subcortical segmentation. Despite some recent efforts to predict it automatically (Klapwijk et al., 2019; Robinson et al., 2019), the assessment of their performance is still relying mostly on visual inspection. Pre-rendering summarized representations of these results, or *snapshots*, instead of any manual procedure involving standard visualization software (e.g., freeview, fsleyes, BrainVisa/Anatomist, and mricron) is a way to minimize time costs and risks of errors. To ease their generation from any Python-enabled environment, we released *nisnap* (Operto and Huguet, 2020). Through one main *plot_segment()* function, it includes controls for opacity, layout, color map, plane/slice selection, label picking, static, or animated rendering. Users may also choose between contours or solid color rendering. Though it also features a specific submodule for XNAT integration, *nisnap* is designed to be used with any individual NIfTI images. The function compiles a figure made of a selected set of slices, both from the input segmentation and (if provided) the original image, and renders an overlay of the former over the latter with the desired options. Animated mode generates a GIF animation with a fading effect on the segmentation. Eventually, an image file is created at the specified location with the resulting snapshot.

The tool may be used from Python scripts or command-line interfaces for offscreen rendering or from Jupyter notebooks for real-time visualization. In our context, *Validators* rely on *nisnap* to convert results from SPM, FreeSurfer, or ASHS into snapshots which are then included in validation reports (section "Generalized Automatic Sanity Check/Quality Control"). Snapshots are then collected in a subsequent step for visual review using *snaprate* (section "Assisted Visual QC: *snaprate*"). **Figure 4** shows an example of snapshot produced by *nisnap* and displayed for review through *snaprate*.

nisnap is released as an independent tool, available as a PyPI package and code is hosted on GitHub[7].

Assisted Visual QC: *snaprate*

Automatic controls performed by *Validators* include generation of snapshots (e.g., for segmentation results using SPM, CAT, FreeSurfer, processing of diffusion-weighted imaging data using FSL, and registration using ANTs, among others). Although navigation is not enabled as it would be with a full-featured NIfTI viewer, e.g., Papaya[8], brainbrowser (Sherif et al., 2015), and brainbox (Heuer et al., 2016), snapshots are lightweight

[7]http://github.com/xgrg/nisnap

[8]https://github.com/rii-mango/Papaya

resources that are displayed instantly and easily cacheable at runtime, hence resulting in optimized overall time of review. Such rendered representations allow fine-grained customization and are suited for the review of large collections of data. Nevertheless, they can still not be checked in a fully automatic way and generally require visual inspection. In particular, such an approach involving tool-assisted visual review of summarized versions of processing results has already been proposed, e.g., based on MR slices (Raamana, 2018) or pre-generated snapshots (Keshavan et al., 2019). Some alternatives include features for real-time NIfTI visualization and manual voxel labeling, thus enabling crowdsourced reviews and corrections (Heuer et al., 2016; Keshavan et al., 2018).

In line with this–and in order to minimize the burden given to experts and optimize the review process–we present an assisting tool (**Figure 4**) that naturally connects to the previously described system, collects previously generated snapshots (along with an optional predefined set of useful *Test* outputs), and displays them within a multi-user collaborative web application. Registered raters may navigate and assign each of them with a descriptive comment and a quality score. Snapshots are produced prior to the review process during automatic individual report generation, described in the previous section. Rendering is done based on either *nilearn.plotting* submodule (Abraham et al., 2014) or *nisnap* (as described in section "Generating Summarized Representations of Segmentation Results: *nisnap*").

As snapshots are generated during the execution of Validators and their corresponding Tests, they may then be displayed along with the outcomes from those prior checkpoints. For instance, segmentation results produced by SPM12 come with prior Tests such as HasNormalVolumes ("do global gray/white matter volumes fall inside predefined target intervals?") or SPM12SegmentExecutionTime ("did the pipeline take longer than a given threshold?"). Such checkpoints may be displayed under the snapshot to provide additional assistance to the review process. One of them can be selected, at the user's choice, so that the navigation will jump from one failed case to the following one. In case further inspection of a given case is required, a direct link takes the user to the corresponding experiment on the XNAT platform. Users are also allowed to switch between pipelines/types of snapshots to assess their quality over the same subject (**Figure 4**).

We present *snaprate* (Operto, 2019) in its particular XNAT-centric software ecosystem. Nevertheless, the tool itself is designed to work alone with any type of pre-generated snapshots or figures. Here, image-based processing outputs are represented as a collection of slices either from the original images (e.g., fractional anisotropy or tensor maps from FSL DTIFIT) or from the original T1-weighted images overlaid with the segmentation/registration results (e.g., from SPM, CAT, FreeSurfer, ASHS, ANTs) (**Figure 4**). Prior to the review, all snapshots are extracted from reports and bulk downloaded into a single folder using *bx*[9]. Then, *snaprate* operates as a web application (using the *Tornado*[10] Python web framework) on which users may log in using their individual browser. Every action (addition/edit of any score/comment) is automatically stored server-side as tabular data and may also be downloaded locally as spreadsheet files.

Code is available on GitHub at: http://github.com/xgrg/snaprate and a full demo can be found at http://snaprate.herokuapp.com.

DISCUSSION

Recent decades have witnessed an increasing number of large to very large imaging studies, prominently in the field of neurodegenerative diseases. The datasets collected during these studies form essential resources for the research aiming at new biomarkers. Nevertheless, setting up a basic infrastructure to collect, host, manage, process, review, and share those datasets is still a hard task, especially for organizations with their own imaging equipment, and the number of options in terms of existing open-source software platforms for neuroinformatics facilitating the seamless connection of an imaging scanner is still quite limited. Larger projects may afford to develop their own systems to serve these datasets, hence providing high-performance and customized service (e.g., primary access to the data, to computational resources, algorithms) to a restricted set of users. However, such systems are rarely designed to provide reusable solutions that could be easily adapted elsewhere. As opposed to this, the approach described in this article is characterized by its low footprint and high modularity, hence facilitating selective reuse and allowing incremental development. By low footprint, we suggest that the presented components not only introduce little dependencies (i.e., essential Python libraries) but also work with basic human-friendly objects (e.g., spreadsheets, JSON files, JPEG images, and PDF documents) making them again easily reusable independently.

The approach was implemented and is currently running in the context of an individual research institution managing cohort programs on risk factors and biomarkers of AD: the BBRC. It may in itself serve as a practical example for organizations with similar purposes. Such an empirical description, though, may not substitute a proper comparative study, not presented in this article, to assess the relative performance of this model. Nevertheless, it was built following guiding principles taken from best coding practices and software quality (e.g., extensibility, reusability, minimum cost to develop, clear definition of purpose) (Hoare, 1972). In that regard, all described components (*bx*, *nisnap*, *snaprate*, *bbrc-validator*) include diligent automated testing for CI (e.g., through sandboxed executions of most commands), thus yielding code coverage rates consistently over 90%. Additionally, as described in section "Generalized Automatic Sanity Check/Quality Control," each *Test* in every *Validator* is, by definition, assigned with two imaging sessions, one that is expected to pass and the other, to fail. This not only complements the *Test*'s documentation by providing the reader with genuine examples but also ensures that *Tests* are systematically *tested* against real-life cases after every new change in the code. It is also worth noting that those current

[9]Using the command: `bx snapshot <project>`.
[10]http://www.tornadoweb.org

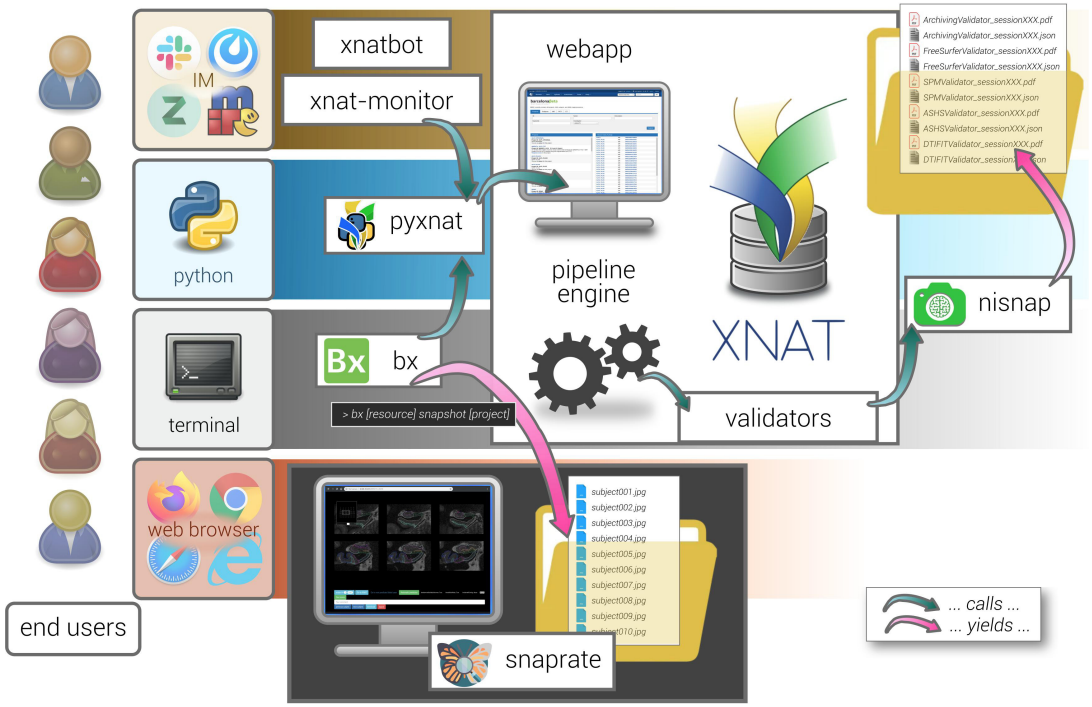


FIGURE 5 | General view on the XNAT-based ecosystem architecture. The different satellite tools described in this manuscript are represented with their mutual interactions. Each of them is based on a specific type of user interaction, e.g., command line, scripts, web browser, and instant messaging (IM). Interaction with XNAT (e.g., xnatbot, xnat-monitor, bx, nisnap) relies on the pyxnat library. *Validators* are run as pipelines and produce reports (calling nisnap for snapshot generation). Snapshots are collected from any given XNAT project thanks to one of bx's commands (*snapshot*) and passed to snaprate for visual quality assessment.

Validators (as the ones featured in **Supplementary Table 1**) have been tailored to the needs of one specific organization (e.g., checking the software version of a Philips MR scanner) and may be considered neither comprehensive nor suited for other institutions. However, the modularity and flexibility of the system allow them to easily adapt them to their respective contexts.

Another potential limitation of this present model is that by mostly focusing on automatic outputs, it is not well-adapted to handle manual corrections. In this version, workflows are automatically launched and managed through the XNAT Pipeline Engine, and their history is stored and searchable in the XNAT database. Pipelines are defined by a set of dependencies and conditions based on other pipelines and prior automatic tests. Failing cases are then flagged and ignored in subsequent steps. One drawback of this conservative approach is that failed cases (failed workflows or QC) are simply discarded from further analysis, resulting currently in a line loss of data that could probably be harnessed if processed manually. On the other hand, this strategy, by limiting manually input data/parameters, avoids the creation of forks and makes traceability easier to control by guaranteeing that any resource can only have a linear history. In this respect, coupling the system to a solution like DataLad (Wagner et al., 2019) to address version control may provide an interesting avenue for improvement.

The overall system is built around XNAT, which is among the most broadly deployed open source systems for managing medical imaging data in research (Nichols and Pohl, 2015). We then enriched the platform with QC-oriented features by taking advantage of its REST API using Python (Schwartz et al., 2012). QC is balanced between automatic tests and tool-assisted visual inspection. On the one hand, automatic operations include sanity checks, collection of quality metrics, quality prediction, and generation of human-readable reports, all part of a single module, *bbrc-validator*, which was designed to have new tests easily added (and covered by CI automated testing). On the other hand, visual inspection is based on collaborative review of pre-rendered snapshots. **Figure 5** illustrates this XNAT-centered ecosystem as a whole.

With all neuroimaging studies growing in scale and complexity, QA/QC has become a difficult task and a heavy burden, which is managed in very heterogeneous ways across research groups (depending on data type, sample size, experience, and resource availability, among others). Emergence of standardized QC methods is still required and is currently hindered by the existing variety of acquisition protocols (modalities and scanner manufacturers) or processing pipelines. While efforts have been initiated by the community in this regard–e.g., by the INCF Special Interest Group on Neuroimaging Quality Control[11] (niQC)–common frameworks remain needed to make QC-related tasks easier and more efficient, with enough practical flexibility to be adapted across

[11]https://incf.github.io/niQC/

different contexts, and hence contribute to ongoing discussions on standardization. Mistakes and errors are inevitable: such a model as the one described in this paper does not claim to eradicate them all, but to reduce their likelihood and severity by punctuating workflows with tailored checkpoints and safeguards. New caught inconsistencies get converted into new control points, increasing general "test coverage rate" (Miller and Maloney, 1963) across iterations, hence tending toward better global data quality assessment in the long run–provided no changes affect the data source.

We also think that such a model, by integrating a routine automatic collection of quality-related parameters, on one side, and a component for facilitated collaborative visual review, on the other, may efficiently serve as a stepping stone for improved automatic classifiers for QC and potentially contribute with new crowdsourced quality metrics, as proposed by Esteban et al. (2019a). Following this, one interesting future development would be to connect *snaprate* to MRIQC's automatic prediction (Esteban et al., 2017).

On a different level, tools like *monitors* or *bx* are also based on XNAT, through calls to its REST API using *pyxnat*, and as such help in achieving customized and diversified user experience with the database.

We hence present a collection of basic individual components that, taken as a whole, form a novel ecological arrangement based on strong core principles (lightweight, reuse of existing tools, and reproducibility), which has shown efficiency in the context of single-site imaging cohort studies conducted by an individual research platform. Again, modularity makes it easy to take one or several components and allow their reuse by other groups, primarily the ones making use of large neuroimaging datasets for their research.

Finally, some of the presented components such as *snaprate* or *nisnap* are purely independent from XNAT since they are based on source-agnostic snapshots and as such may be used in any framework. The other ones are interfaced with the platform core using a unique library, *pyxnat* (**Figure 5**), therefore making the whole system virtually adaptable to other types of platforms just by replacing the binding module. Nevertheless, by leveraging its built-in features in particular for access right management, we believe that having XNAT as a cornerstone of the model is bound to have a downstream positive impact on data sharing (Herrick et al., 2016), primarily in groups lacking the necessary technical support (Poline et al., 2012; Haselgrove et al., 2014).

CONCLUSION

Quality control of neuroimaging datasets and their processed derivatives is still an open problem in all cohort studies and generally synonymous with heavy burden. Its strong dependence on protocol specifications (i.e., study design, imaging protocol, and processing workflows) hinders the adoption of standardized approaches. Furthermore, the nature of subsequent analyses is also linked to the right verification procedure to implement and the same dataset may have to go through different QC passes depending on the final research question. To cope with this, a substantial amount of intermediate control steps may be automatized, as described in this paper, while the remaining needed visual inspection may be facilitated by integrated collaborative semi-automatic tools. As both aspects are tightly interconnected, all these QC procedures must be supported by some flexible and efficient data management strategies. We showed in that context that, capitalizing on existing components and by only adding some light interaction layer between them, user experience in accessing data can be diversified and thus fit with a variety of user profiles. Hence, providing improved access to data at its source is bound to give way to better analysis workflows in terms of traceability and reproducibility. All these components take part in a whole ecosystem that has been assembled and is currently running at the BBRC, an individual research unit managing cohort research programs on AD. By its modularity and the lightweight footprint/reusability of its parts, this ecosystem may be easily adjusted and/or augmented in accordance with other research groups' needs.

MEMBERS OF THE ALFA STUDY

The following are the collaborators of the ALFA study: Müge Akinci, Annabella Beteta, Raffaele Cacciaglia, Alba Cañas, Irene Cumplido, Carme Deulofeu, Ruth Dominguez, Maria Emilio, Karine Fauria, Sherezade Fuentes, Oriol Grau-Rivera, José M. González de Echevarri, Laura Hernandez, Gema Huesa, Iva Knezevic, Eider M. Arenaza-Urquijo, Eva M. Palacios, Paula Marne, Marta Milà-Alomà, Tania Menchón, Carolina Minguillon, Albina Polo, Sandra Pradas, Blanca Rodríguez, Aleix Sala Vila, Gemma Salvadó, Gonzalo Sánchez-Benavides, Mahnaz Shekari, Anna Soteras, Laura Stankeviciute, Marc Suárez-Calvet, Marc Vilanova, and Natalia Vilor-Tejedor.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The ALFA study was approved by the Independent Ethics Committee "Parc de Salut Mar," Barcelona. All participating subjects and signed the study's informed consent form that had also been approved by the Independent Ethics Committee "Parc de Salut Mar", Barcelona. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JH co-implemented the ecosystem and co-wrote the manuscript. CF contributed to data acquisition and analysis and provided a critical revision of the manuscript. DF provided a critical revision

of the manuscript. SG and DV provided expertise in high-performance computing and critical revision of the manuscript. JM and JG supported the design of the whole study, provided a critical revision of the manuscript, and supervised this project. GO designed and co-implemented the tools, supervised the development of the ecosystem, and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

The research leading to these results has received funding from "la Caixa" Foundation (ID 100010434), under agreement LCF/PR/GN17/50300004. Additional funding was obtained from the Universities and Research Secretariat, Ministry of Business and Knowledge of the Catalan In review 16 Government under grant no. 2017-SGR-892. JG is supported by the Spanish Ministry of Economy and Competitiveness (RYC-2013-13054). JG has received research support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking AMYPAD grant agreement no. 115952 and from Ministerio de Ciencia, Innovación y Universidades (grant agreement RTI2018-102261).

ACKNOWLEDGMENTS

This publication is part of the ALFA study (Alzheimer and Families). The authors would like to express their most sincere gratitude to the ALFA project participants, without whom this research would have not been possible. The authors would also like to warmly thank Dr. González de Echavarri for their artistic contribution to this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.633438/full#supplementary-material

REFERENCES

Abe, S., Irimia, A., and Van Horn, J. D. (2015). Quality control considerations for the effective integration of neuroimaging data. *Lecture Notes Comput. Sci.* 9162, 195–201. doi: 10.1007/978-3-319-21843-4_15

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinformat.* 8:14. doi: 10.3389/fninf.2014.00014

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034

Aryanto, K. Y., Broekema, A., Oudkerk, M., and Van Ooijen, P. (2012). Implementation of an anonymisation tool for clinical trials using a clinical trial processor integrated with an existing trial patient data information system. *Eur. Radiol.* 22, 144–151. doi: 10.1007/s00330-011-2235-y

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *NeuroImage* 26, 839–851. doi: 10.1016/J.NEUROIMAGE.2005.02.018

Avants, B., Tustison, N., and Song, G. (2009). *Advanced Normalization Tools (ANTS). Insight Journal*. Available online at: http://hdl.handle.net/10380/3113.

Avesani, P., McPherson, B., Hayashi, S., Caiafa, C. F., Henschel, R., Garyfallidis, E., et al. (2019). The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *Scie. Data* 6:69. doi: 10.1038/s41597-019-0073-y

Bastiani, M., Cottaar, M., Fitzgibbon, S. P., Suri, S., Alfaro-Almagro, F., Sotiropoulos, S. N., et al. (2019). Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction. *NeuroImage* 184, 801–812. doi: 10.1016/j.neuroimage.2018.09.073

Book, G. A., Anderson, B. M., Stevens, M. C., Glahn, D. C., Assaf, M., and Pearlson, G. D. (2013). Neuroinformatics database (n.d.) - A modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics* 11, 495–505. doi: 10.1007/s12021-013-9194-1

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., et al. (2018). The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* 32, 43–54. doi: 10.1016/j.dcn.2018.03.001

Cox, S., Ritchie, S., Fawns-Ritchie, C., Tucker-Drob, E., and Deary, I. (2019). Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence* 76:101376.doi:10.1016/j.intell.2019.101376

Das, S., Zijdenbos, A. P., Harlap, J., Vins, D., and Evans, A. C. (2010). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinform.* 4:37. doi: 10.3389/fninf.2011.00037

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One* 12:e0184661. doi: 10.1371/journal.pone.0184661

Esteban, O., Blair, R. W., Nielson, D. M., Varada, J. C., Marrett, S., Thomas, A. G., et al. (2019a). Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines. *Scie. Data* 6:30. doi: 10.1038/s41597-019-0035-4

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019b). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116. doi: 10.1038/s41592-018-0235-4

Fischl, B. (2012). FreeSurfer. *NeuroImage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Frisoni, G. B., Redolfi, A., Manset, D., Rousseau, M. -É, Toga, A., and Evans, A. C. (2011). Virtual imaging laboratories for marker discovery in neurodegenerative diseases. *Nat. Rev. Neurol.* 7, 429–438. doi: 10.1038/nrneurol.2011.99

Gaser, C. (2016). *CAT - A Computational Anatomy Toolbox for SPM*. Minnesota: Organization for Human Brain Mapping.

Gee, T., Kenny, S., Price, C. J., Seghier, M. L., Small, S. L., Leff, A. P., et al. (2010). Data warehousing methods and processing infrastructure for brain recovery research. *Arch. Ital. Biol.* 148, 207–217.

Glatard, T., Kiar, G., Aumentado-Armstrong, T., Beck, N., Bellec, P., Bernard, R., et al. (2018). Boutiques: a flexible framework to integrate command-line applications in computing platforms. *GigaScience* 7:giy016. doi: 10.1093/gigascience/giy016

Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B., and Poldrack, R. (2017). OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. *Organ. Hum. Brain Mapp. Vancouver Canada* 6:1677. doi: 10.1038/sdata.2016.44.3

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44

Gutman, D. A., Dunn, W. D., Cobb, J., Stoner, R. M., Kalpathy-Cramer, J., and Erickson, B. (2014). Web based tools for visualizing imaging data and development of XNATView, a zero footprint image viewer. *Front. Neuroinformat.* 8:53. doi: 10.3389/fninf.2014.00053

Harrigan, R. L., Yvernault, B. C., Boyd, B. D., Damon, S. M., Gibney, K. D., Conrad, B. N., et al. (2016). Vanderbilt university institute of imaging science center for computational imaging XNAT: a multimodal data archive and processing environment. *NeuroImage* 124, 1097–1101. doi: 10.1016/j.neuroimage.2015.05.021

Haselgrove, C., Poline, J.-B., and Kennedy, D. N. (2014). A simple tool for neuroimaging data sharing. *Front. Neuroinform.* 8:52. doi: 10.3389/fninf.2014.00052

Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., and Marcus, D. S. (2016). XNAT central: open sourcing imaging research data. *NeuroImage* 124, 1093–1096. doi: 10.1016/j.neuroimage.2015.06.076

Heuer, K., Ghosh, S., Robinson Sterling, A., and Toro, R. (2016). Open neuroimaging laboratory. *Res. Ideas Outcomes* 2:e9113. doi: 10.3897/rio.2.e9113

Heunis, S., Hellrung, L., Van der Meer, B. S., Sladky, R., Pamplona, G. S., Scharnowski, F., et al. (2019). *rtQC: an Open-Source Toolbox for Real-Time fMRI Quality Control*. Minnesota: Organization for Human Brain Mapping, doi: 10.5281/zenodo.3239084

Hoare, C. A. (1972). The quality of software. *Software* 2, 103–105. doi: 10.1002/spe.4380020202

Ito, K. L., Kumar, A., Zavaliangos-Petropulu, A., Cramer, S. C., and Liew, S.-L. (2018). Pipeline for analyzing lesions after stroke (PALS). *Front. Neuroinform.* 12:63. doi: 10.3389/fninf.2018.00063

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Job, D. E., Dickie, D. A., Rodriguez, D., Robson, A., Danso, S., Pernet, C., et al. (2017). A brain imaging repository of normal structural MRI across the life course: brain images of normal subjects (BRAINS). *NeuroImage* 144, 299–304. doi: 10.1016/j.neuroimage.2016.01.027

Johnson, H. A. (2018). Slack. *J. Med. Libr. Assoc.* 106, 148–151. doi: 10.5195/JMLA.2018.315

Júlvez, J., Paus, T., Bellinger, D., Eskenazi, B., Tiemeier, H., Pearce, N., et al. (2016). Environment and brain development: challenges in the global context. *Neuroepidemiology* 46, 79–82. doi: 10.1159/000442256

Keator, D., Helmer, K., Steffener, J., Turner, J., Van Erp, T., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *NeuroImage* 82, 647–661. doi: 10.1016/J.NEUROIMAGE.2013.05.094

Kennedy, D. N., Abraham, S. A., Bates, J. F., Crowley, A., Ghosh, S., Gillespie, T., et al. (2019). Everything matters: the repronim perspective on reproducible neuroimaging. *Front. Neuroinform.* 13:1. doi: 10.3389/fninf.2019.00001

Kennedy, D. N., Haselgrove, C., Riehl, J., Preuss, N., and Buccigrossi, R. (2016). The NITRC image repository. *NeuroImage* 124, 1069–1073. doi: 10.1016/j.neuroimage.2015.05.074

Keshavan, A., Datta, E., McDonough, I. M., Madan, C. R., Jordan, K., and Henry, R. G. (2018). Mindcontrol: a web application for brain segmentation quality control. *NeuroImage* 170, 365–372. doi: 10.1016/j.neuroimage.2017.03.055

Keshavan, A., Yeatman, J. D., and Rokem, A. (2019). Combining citizen science and deep learning to amplify expertise in neuroimaging. *Front. Neuroinform.* 13:29. doi: 10.3389/fninf.2019.00029

Kiar, G., Gorgolewski, K. J., Kleissas, D., Roncal, W. G., Litt, B., Wandell, B., et al. (2017). Science in the cloud (SIC): a use case in MRI connectomics. *GigaScience* 6:gix013. doi: 10.1093/gigascience/gix013

Klapwijk, E. T., van de Kamp, F., van der Meulen, M., Peters, S., Wierenga, L. M., van de Kamp, F., et al. (2019). Qoala-T: a supervised-learning tool for quality control of FreeSurfer segmented MRI data. *NeuroImage* 189, 116–129. doi: 10.1016/j.neuroimage.2019.01.014

Leemans, A., Jeurissen, B., Sijbers, J., and Jones, D. K. (2009). ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. *Proc. Int. Soc. Mag. Reson. Med.* 17:3537.

Manjón, J. V., and Coupé, P. (2016). Volbrain: an online MRI brain volumetry system. *Front. Neuroinform.* 10:30. doi: 10.3389/fninf.2016.00030

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* 22, 2677–2684. doi: 10.1162/jocn.2009.21407

Marcus, D. S., Harwell, J., Olsen, T., Hodge, M., Glasser, M. F., Prior, F., et al. (2011). Informatics and data mining tools and strategies for the human connectome project. *Front. Neuroinform.* 5:4. doi: 10.3389/fninf.2011.00004

Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit. *Neuroinformatics* 5, 11–33. doi: 10.1385/ni:5:1:11

Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). Making data sharing work: the FCP/INDI experience. *NeuroImage* 82, 683–691. doi: 10.1016/j.neuroimage.2012.10.064

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* 25, 1–20. doi: 10.1097/01.NND.0000320699.47006.a3

Miller, J. C., and Maloney, C. J. (1963). Systematic mistake analysis of digital computer programs. *Commun. ACM* 6, 58–63. doi: 10.1145/366246.366248

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536. doi: 10.1038/nn.4393

Molinuevo, J. L., Gramunt, N., Gispert, J. D., Fauria, K., Esteller, M., Minguillon, C., et al. (2016). The ALFA project: a research platform to identify early pathophysiological features of Alzheimer's disease. *Alzheimer's Dement.* 2, 82–92. doi: 10.1016/j.trci.2016.02.003

Mutsaerts, H., Petr, J., Lysvik, E., Schrantee, A., Shirzadi, Z., Zelaya, F., et al. (2017). "ExploreASL: image processing toolbox for multi-center arterial spin labeling population analyses," in *34th Annual Scientific Meeting of European Society for Magnetic Resonance in Medicine and Biology*, Barcelona.

Nichols, B. N., and Pohl, K. M. (2015). Neuroinformatics software applications supporting electronic data capture, management, and sharing for the neuroimaging community. *Neuropsychol. Rev.* 25, 356–368. doi: 10.1007/s11065-015-9293-x

Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., et al. (2014). DTIPrep: Quality control of diffusion-weighted images. *Front. Neuroinform.* 8:4. doi: 10.3389/fninf.2014.00004

Operto, G. (2019). *snaprate v1.0*. Geneva: Zenodo, doi: 10.5281/zenodo.3539331

Operto, G., and Huguet, J. (2020). *nisnap v0.3.7*. Geneva: Zenodo, doi: 10.5281/zenodo.4075418

Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012

Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K. F., Halchenko, Y. O., and Hanke, M. (2012). Data sharing in neuroimaging research. *Front. Neuroinform.* 6:9. doi: 10.3389/fninf.2012.00009

Raamana, P. R. (2018). *VisualQC: Assistive Tools for Easy and Rigorous Quality Control of Neuroimaging Data (Version 0.28)*. Genèv: Zenodo, doi: 10.5281/zenodo.1211365

Raymond, E. S. (1999). The cathedral and the bazaar. *Knowledge, Technology and Policy* 12, 23–49. doi: 10.1007/s12130-999-1026-0

Redolfi, A., Manset, D., Barkhof, F., Wahlund, L.-O., Glatard, T., Mangin, J.-F., et al. (2015). Head-to-head comparison of two popular cortical thickness extraction algorithms: a cross-sectional and longitudinal study. *PLoS One* 10:e0117692. doi: 10.1371/journal.pone.0117692

Ritchie, C. W., Muniz-Terrera, G., Kivipelto, M., Solomon, A., Tom, B., and Molinuevo, J. L. (2020). The european prevention of Alzheimer's Dementia (EPAD) longitudinal cohort study: baseline data release V500.0. *J. Prev. of Alzheimer's Dis.* 7, 8–13. doi: 10.14283/jpad.2019.46

Robinson, R., Valindria, V. V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., et al. (2019). Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. *J. Cardiovasc. Magn. Reson.* 21:18. doi: 10.1186/s12968-019-0523-x

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., et al. (2010). The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry* 15, 1128–1139. doi: 10.1038/mp.2010.4

Schwartz, Y., Barbot, A., Thyreau, B., Frouin, V., Varoquaux, G., Siram, A., et al. (2012). PyXNAT: XNAT in python. *Front. Neuroinform.* 6:12. doi: 10.3389/fninf.2012.00012

Scott, C. J., Arnott, S. R., Chemparathy, A., Dong, F., Solovey, I., Gee, T., et al. (2020). An overview of the quality assurance and quality control of magnetic resonance imaging data for the ontario neurodegenerative disease research initiative (ONDRI): Pipeline development and neuroinformatics. *bioRxiv* [Preprint]. doi: 10.1101/2020.01.10.896415

Shehzad, Z., Giavasis, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., et al. (2015). The preprocessed connectomes project quality assessment protocol - a resource for measuring the quality of MRI data. *Front. Neurosci.* 9:47. doi: 10.3389/conf.fnins.2015.91.00047

Shenkin, S. D., Pernet, C., Nichols, T. E., Poline, J. B., Matthews, P. M., van der Lugt, A., et al. (2017). Improving data availability for brain image biobanking in healthy subjects: Practice-based suggestions from an international multidisciplinary working group. *NeuroImage* 153, 399–409. doi: 10.1016/j.neuroimage.2017.02.030

Sherif, T., Kassis, N., Rousseau, M.-E., Adalat, R., and Evans, A. C. (2015). BrainBrowser: distributed, web-based neurological data visualization. *Front. Neuroinform.* 8:89. doi: 10.3389/fninf.2014.00089

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., et al. (2019). The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project. *BMC Med. Res. Methodol.* 19:102. doi: 10.1186/s12874-019-0737-5

Van Horn, J. D., and Toga, A. W. (2014). Human neuroimaging as a "Big Data" science. *Brain Imaging Behav.* 8, 323–331. doi: 10.1007/s11682-013-9255-y

Wagner, A. S., Waite, L. K., Meyer, K., Heckner, K. M., Kadelka, T., Reuter, N. Y., et al. (2019). *The DataLad Handbook*. Geneva: Zenodo, doi: 10.5281/zenodo.3608612

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Yeatman, J. D., Richie-Halford, A., Smith, J. K., Keshavan, A., and Rokem, A. (2018). A browser-based tool for visualization and analysis of diffusion MRI data. *Nat. Commun.* 9:940. doi: 10.1038/s41467-018-03297-7

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S. L., Gertje, E. C., et al. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Hum. Brain Mapp.* 36, 258–287. doi: 10.1002/hbm.22627

Conflict of Interest: JM is currently a full-time employee of Lundbeck and priorly has served as a consultant or at advisory boards for the following for-profit companies, or has given lectures in symposia sponsored by the following for-profit companies: Roche Diagnostics, Genentech, Novartis, Lundbeck, Oryzon, Biogen, Lilly, Janssen, Green Valley, MSD, Eisai, Alector, BioCross, GE Healthcare, and ProMIS Neurosciences. JG has received speaker's fees from Biogen and Philips.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 29 April 2021
doi: 10.3389/fnins.2021.608808

Transductive Transfer Learning for Domain Adaptation in Brain Magnetic Resonance Image Segmentation

Kaisar Kushibar[1], Mostafa Salem[1,2], Sergi Valverde[1], Àlex Rovira[3], Joaquim Salvi[1], Arnau Oliver[1] and Xavier Lladó[1]*

[1] Institute of Computer Vision and Robotics, University of Girona, Girona, Spain, [2] Computer Science Department, Faculty of Computers and Information, Assiut University, Asyut, Egypt, [3] Magnetic Resonance Unit, Department of Radiology, Vall d'Hebron University Hospital, Barcelona, Spain

Segmentation of brain images from Magnetic Resonance Images (MRI) is an indispensable step in clinical practice. Morphological changes of sub-cortical brain structures and quantification of brain lesions are considered biomarkers of neurological and neurodegenerative disorders and used for diagnosis, treatment planning, and monitoring disease progression. In recent years, deep learning methods showed an outstanding performance in medical image segmentation. However, these methods suffer from generalisability problem due to inter-centre and inter-scanner variabilities of the MRI images. The main objective of the study is to develop an automated deep learning segmentation approach that is accurate and robust to the variabilities in scanner and acquisition protocols. In this paper, we propose a transductive transfer learning approach for domain adaptation to reduce the domain-shift effect in brain MRI segmentation. The transductive scenario assumes that there are sets of images from two different domains: (1) source—images with manually annotated labels; and (2) target—images without expert annotations. Then, the network is jointly optimised integrating both source and target images into the transductive training process to segment the regions of interest and to minimise the domain-shift effect. We proposed to use a histogram loss in the feature level to carry out the latter optimisation problem. In order to demonstrate the benefit of the proposed approach, the method has been tested in two different brain MRI image segmentation problems using multi-centre and multi-scanner databases for: (1) sub-cortical brain structure segmentation; and (2) white matter hyperintensities segmentation. The experiments showed that the segmentation performance of a pre-trained model could be significantly improved by up to 10%. For the first segmentation problem it was possible to achieve a maximum improvement from 0.680 to 0.799 in average Dice Similarity Coefficient (DSC) metric and for the second problem the average DSC improved from 0.504 to 0.602. Moreover, the improvements after domain adaptation were on par or showed better performance compared to the commonly used traditional unsupervised segmentation methods (FIRST and LST), also achieving faster execution time. Taking this into account, this work presents one more step toward the practical implementation of deep learning algorithms into the clinical routine.

Keywords: deep learning, domain adaptation, magnetic resonance imaging, brain, segmentation, sub-cortical structures, white matter hyperintensities, transductive learning

OPEN ACCESS

Edited by:
Diana M. Sima,
Icometrix, Belgium

Reviewed by:
Hongwei Li,
Technical University of Munich, Germany
Emanuele Olivetti,
Bruno Kessler Foundation, Italy
Hristina Uzunova,
University of Lübeck, Germany

****Correspondence:***
Kaisar Kushibar
k.kushibar@gmail.com

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *21 September 2020*
Accepted: *26 March 2021*
Published: *29 April 2021*

Citation:
Kushibar K, Salem M, Valverde S, Rovira À, Salvi J, Oliver A and Lladó X (2021) Transductive Transfer Learning for Domain Adaptation in Brain Magnetic Resonance Image Segmentation. Front. Neurosci. 15:608808. doi: 10.3389/fnins.2021.608808

1. INTRODUCTION

Medical image segmentation is a pivotal task in diagnosis, treatment, and surgical planning, and monitoring disease progression over time. Quantification of brain structures and brain lesions from Magnetic Resonance Images (MRI) is crucial as they are biomarkers for neurological and neurodegenerative disorders. However, manually annotating MRI images is a time-consuming and a laborious task, which has to be done by experts with knowledge in disease-specific aspects and anatomy. Therefore, there is a need for accurate and automated methods to carry out different segmentation problems in brain MRI—e.g., brain structure (González-Villà et al., 2016), multiple sclerosis (MS) (García-Lorenzo et al., 2013), and brain tumour (Bakas et al., 2018).

In recent years, deep learning methods—in particular, Convolutional Neural Networks (CNNs)—have shown a remarkable advance in the field of brain MRI segmentation for many different applications (Akkus et al., 2017; Bernal et al., 2019). Unlike the traditional hand-crafted features, CNNs learn task-specific features directly from observed data (LeCun et al., 2015). Most CNN based approaches for medical image segmentation in literature are usually trained and tested with images that share common characteristics—the same scanner and acquisition protocol. However, the performance of such pre-trained networks decline when tested on images with different MRI characteristics, i.e., images from a different domain (MRI scanner, protocol). Deep learning methods cannot generalise to unseen domains where the image scans vary in brightness, contrast, and resolution. Therefore, the network has to be re-trained using the images from this new domain, requiring expert annotated labels. This commonly faced issue is known as the domain-shift problem, which hinders the applicability of deep learning methods in practice. Moreover, the data-driven nature, which demands a vast amount of expert annotated images, often makes fully retraining a CNN impossible.

Transfer learning strategy is an effective way to adapt a pre-trained neural network to a new domain. This procedure consists in retraining only a few last layers, which can be done using a remarkably smaller number of annotated images (Ghafoorian et al., 2017; Valverde et al., 2019). However, it is not always possible to obtain even a few images to perform transfer learning for domain adaptation. Therefore, other unsupervised domain adaptation methods are active research topics in medical image analysis. A recent work of Orbes-Arteainst et al. (2019) proposed an unsupervised domain adaptation approach in a similar fashion to transfer learning with teacher-student learning strategy. The authors used knowledge-distillation technique where a supervised teacher model is used to train a student network by generating soft labels for the target domain.

In general, unsupervised domain adaptation methods could be categorised into: (1) image-level, where the images of two domains are harmonised to share similar characteristics; and (2) feature-level approaches where the CNN itself is adapted to be more invariant to different imaging domains. Common approaches for the image-level domain adaptation include traditional pre-processing steps (Shah et al., 2011; Fortin et al., 2016). One of the common challenges of the traditional approaches include image artefacts that may appear during intensity transformations that reduce the image quality. Moreover, it was shown (Kushibar et al., 2019) that approaches such as standardising images using the Nyúl histogram matching (Nyúl et al., 2000) or mixing datasets from different domains during training cannot overcome the effect of the domain-shift.

More complex Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) based approaches have also been introduced for translating images into a new target domain. However, most of the works in the literature propose synthesising images from a different imaging modality. For example, Huo et al. (2018), utilise CycleGAN framework to generate CT images from MRI to allow splenomegaly segmentation without using manual annotation on CT. Also, Zhang et al. (2018) proposed a modified CycleGAN approach for multi-organ segmentation on X-ray images using Digitally Reconstructed Radiographs by performing pixel-to-pixel style transfer from one modality to another. Although such approaches have shown promising results, there is still a lack of GAN based methods for single-modality image harmonisation.

Some feature-level domain adaptation methods have also been proposed in recent years. Such methods employ a transductive learning strategy for domain adaptation. In the transductive scenario, the images without expert annotations from unseen domain are included in the training process with the aim to minimise the domain-shift effect. Adversarial training of the network is a well-known transductive learning method. Similarly to GAN architectures, the training strategy consists of two network paths: one for classifying the input patch, and another to force the network to learn domain-invariant features by discriminating source and target domains. Recent work of Kamnitsas et al. (2017) utilises an adversarial training approach for unsupervised domain adaptation from Gradient Echo images to Susceptibility Weighted Images for brain lesion segmentation task. Moreover, an adversarial domain adaptation from Whole Slide pathology to Microscopy images has been studied in Zhang et al. (2019). Chen et al. (2020) proposed simultaneous image to image translation and domain alignment between CT and MRI images using a modification of a CycleGAN for cardiac and abdominal multi-organ segmentation. However, more investigation is needed for the adversarial training for domain adaptation for a scenario where the domain difference is subtle—i.e., multi-site and single-modality images.

There are some drawbacks of GAN based and adversarial training strategies. These methods are usually formulated as a competition between two agents: discriminator and segmenter (Yi et al., 2019). In general, the objective for the latter can vary according to the task (e.g., it is called generator for image synthesis), but in most cases the objective of the former is to differentiate between two distributions. In this non-convex min-max formulation, the training of the network can be difficult and unstable, which requires a careful selection of architecture, weight initialisation, and hyper-parameter tuning (Roth et al., 2017). For example, Li et al. (2020) proposed an adversarial approach for single modality domain adaptation with flip-label technique

where the labels of the discriminator model were partly inverted during training to minimise over-fitting.

Other feature-level transductive domain adaptation methods perform domain distribution discrepancy minimisation to learn domain-invariant features. Most of the advancements of such approaches are done for computer vision with natural images (Damodaran et al., 2018; Rozantsev et al., 2018; Kang et al., 2019). However, only a few works have been proposed in medical imaging field for single-modality images. One of the recent domain adaptation approaches is the work of Ackaouy et al. (2020) for multi-site brain multiple sclerosis lesion segmentation. The authors adopted a joint distribution optimal transport framework proposed in Damodaran et al. (2018) to compare the source and target distributions and bring them closer in a feature-level.

In this paper, we propose a feature-level transductive domain adaptation method that can be trained without extensive hyper-parameter tuning. Similarly to Ackaouy et al. (2020), our proposed method aligns the network feature distributions between two different domains by forcing the convolutional and fully connected layers to produce similar activation maps by minimising the histogram distribution differences. The images from a new domain are incorporated within training transductively and do not require expert annotated ground truths. To show its robustness and applicability, we utilise and evaluate our domain adaptation approach for two active brain MRI segmentation problems—brain sub-cortical structure segmentation and brain White Matter Hyperintensities (WMH) segmentation. We compare the performance of our proposal with segmentation results without domain adaptation as well as the unsupervised state-of-the-art approaches for each problem: (1) FIRST (Patenaude et al., 2011) for sub-cortical structure segmentation; and (2) LST (Schmidt and Wink, 2019) for WMH lesion segmentation.

2. DATASETS AND PRE-PROCESSING

We used publicly available and in-house datasets to test the performance of our proposed method for the selected segmentation tasks. Internet Brain Segmentation Repository[1] (IBSR) and Multi-Atlas Labelling Challenge (MICCAI2012) datasets (Landman and Warfield, 2012) were used for the sub-cortical structure segmentation problem. For the WMH segmentation, one dataset comes from an international WMH lesion segmentation challenge (Kuijf et al., 2019), and another from the Vall d'Hebron Hospital Centre (Barcelona, Spain). More information for each dataset is given below.

2.1. Sub-cortical Brain Structure Segmentation

2.1.1. Motivation

The sub-cortical structures are located beneath the cerebral cortex and include the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens. Their deviations in volume over time are considered as biomarkers of neurological diseases such as bipolar disorder (Frazier et al., 2005), Alzheimer's (De Jong et al., 2008), schizophrenia (Rimol et al., 2010), Parkinson's disease (Mak et al., 2014), multiple sclerosis (Houtchens et al., 2007), and are used for pre-operative evaluation and surgical planning (Kikinis et al., 1996), and longitudinal monitoring for disease progression or remission (Storelli et al., 2018). The volumes of the sub-cortical structures differ drastically, in average, 8,500 and $\approx$ 550 mm^3 for largest thalamus and smallest accumbens structures, respectively. This makes the segmentation task more challenging by introducing an unbalanced class problem.

2.1.2. Multi-Atlas Labelling Challenge—MICCAI 2012

The MICCAI 2012 dataset consists of 35 T1-w images in total with 15 training and 20 testing MRI scans. In our experiments, we used the 20 testing set only for testing purposes and they were not included in the training or validation processes in order to follow the rules of the Multi-Atlas Labelling challenge. All T1-w scans have 1 mm^3 isotropic resolution and image dimensions are $256 \times 256 \times 256$ voxels. All images in this dataset were acquired using the same Siemens (1.5 T) MRI scanner. Manually annotated ground truth masks were provided for 134 structures in total, from which 14 classes were extracted for the seven sub-cortical structures corresponding to the left and right hemispheres.

2.1.3. Internet Brain Segmentation Repository—IBSR

The IBSR dataset contains 18 T1-w images in total which are publicly available under the Creative Commons: Attribute license (CC-BY, 2020) as part of the Child and Adolescent Neuro-Development Initiative (CANDI) (Kennedy et al., 2012). The image volumes in this dataset come in three different resolutions—$0.84 \times 0.84 \times 1.5$, $0.94 \times 0.94 \times 1.5$, and $1 \times 1 \times 1.5$ mm^3—and were acquired using two different MRI scanners—GE (1.5 T) and Siemens (1.5 T). Manual annotations for all IBSR images were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and consist of 43 different structures in total (Rohlfing, 2012). For our experiments, we selected the 14 labels corresponding to seven sub-cortical structures with left and right parts separately.

2.2. White Matter Hyperintensity Lesion Segmentation

2.2.1. Motivation

White Matter Hyperintensities are brain lesions that appear bright in T2-weighted and Fluid Attenuated Inversion Recovery (FLAIR) sequences. The presence of the WMH lesions can be from different factors including small vessel disease (Van Norden et al., 2011), multiple sclerosis (Kutzelnigg et al., 2005), stroke or dementia (Debette and Markus, 2010). Monitoring the lesion load and appearance of new lesions is important for diagnosis, longitudinal analysis, and treatment planning (Polman et al., 2011). In contrast to the sub-cortical structure segmentation task, WMH lesions can appear anywhere in the brain within the white matter and can be of different shape and size. Taking into account the importance of lesion load quantification as biomarkers for different neurodegenerative disorders, this task is a relevant and a challenging segmentation problem.

[1] https://www.nitrc.org/projects/ibsr.

2.2.2. White Matter Hyperintensities Segmentation Challenge—WMH 2017

The WMH 2017 dataset provides T1-w and FLAIR scans for 60 patients in total and were acquired from three different sites[2]: (1) UMC Utrecht—3T Philips Achieva with 1 mm^3 isotropic T1-w and $0.96 \times 0.95 \times 3.0$ mm^3 resolution FLAIR sequences; (2) NUHS Singapore—3 T Siemens TrioTim with 1 mm^3 isotropic T1-w and $1.0 \times 1.0 \times 3.0$ mm^3 resolution FLAIR sequences; and (3) VU Amsterdam—3 T GE Signa HDxt with $0.94 \times 0.94 \times 1.0$ mm^3 T1-w and $0.98 \times 0.98 \times 1.2$ mm^3 resolution FLAIR sequences. All T1-w volumes were re-sampled to their corresponding FLAIR images. Ground truth labels for the WMH lesions were manually annotated and peer-reviewed by experts (Kuijf et al., 2019).

2.2.3. In-House Dataset—Vall d'Hebron Hospital, Barcelona (VH)

This dataset contains MRI images for 28 patients with clinically isolated syndrome or early relapsing multiple sclerosis. All MRI scans were acquired in the same 3T Siemens TrioTim scanner that include T1-w and FLAIR images with $1.0 \times 1.0 \times 1.2$ and $0.49 \times 0.49 \times 3.0$ mm^3 resolutions, respectively. Similarly to the WMH 2017 dataset, all T1-w images were re-sampled to their corresponding FLAIR sequences. The WMH lesions were manually annotated and peer-reviewed by experts from the Vall d'Hebron Hospital centre. The MRI volumes were included in this dataset after the patients gave their informed consent which was approved by the Institutional Review Board.

3. METHODS

3.1. CNN Architecture

In this work, to study the domain-shift problem and to evaluate our transductive domain adaptation approach, we took the recent architecture proposed in Kushibar et al. (2018), which achieved state-of-the-art performance for sub-cortical brain structure segmentation. The CNN is shown in **Figure 1** and consists of three paths to process 2D patches of size 32×32. Each path is equipped with five convolution layers, which are followed by a fully connected layer. The outputs of these paths are concatenated together with an additional 15 units corresponding to atlas probabilities for the 14 sub-cortical brain structures and the background. According to Kushibar et al. (2018), incorporation of the atlas probabilities as spatial prior to guide the network significantly improved the performance. For the case of WMH lesion segmentation the number of units for the atlas probabilities is changed to three, which correspond to white matter, grey matter, and cerebro-spinal fluid probabilities. Finally, it is followed by fully connected layers to mine and classify the produced output from the preceding layers. Three 2D patches are extracted for every voxel from the axial, sagittal and coronal views of a 3D volume, making 2.5D patch samples. Next, each orthogonal 2D patch of the 2.5D sample is inputted to the three paths of the CNN. Although full 3D patches contain more surrounding information per voxel, it is more memory-demanding than using 2D patches in voxel-wise segmentation setup. Therefore, employing 2.5D patches is a good trade-off between memory and contextual information for the network (Kushibar et al., 2018).

3.2. Pre-processing

Some commonly used image pre-processing techniques were applied to all of the images in the four datasets. First of all, we non-linearly registered atlas probabilities to the images using the fast free-form deformation method (Modat et al., 2010) that was implemented by the NiftyReg tool[3]. We used the well-known Harvard-Oxford probabilistic atlas (Caviness Jr et al., 1996) distributed with the FSL (v5.0) tool[4]. Note that the number of probabilistic maps for the structure segmentation problem is 14, whereas it is 3 for the WMH lesion segmentation which correspond to the three tissue types. In the next step, we skull-stripped all the MRI volumes—i.e., removed non-brain structures, such as the eyes and skull—using the ROBEX (v1.2) tool (Iglesias et al., 2011). Additionally, we performed bias-field correction to remove intensity inhomogeneities from the images using the FSL-FAST tool. All subject volume intensities were normalised to have a zero mean and unit variance before training and testing the pipeline. Note that the images provided in WMH 2017 Challenge were already bias-field-corrected, co-registered, and the 3D T1-weighted images were aligned (re-sampled) with the FLAIR images by the organisers (Kuijf et al., 2019).

3.3. Initial Training

Before adapting the network to a new domain for a certain task, we assume that the network is pre-trained for the same segmentation problem. Therefore, in this section, we describe how the initial training was done for each segmentation task.

For the sub-cortical structure segmentation problem, we used the same initial training process as described in Kushibar et al. (2018). All samples were extracted from the 14 sub-cortical structures, and the background (negative) samples were selected only from the structure boundaries within a five-voxel margin. Extracting the negative samples in this way allows the network to learn the most difficult areas of the region of interest that correspond to the structure borders. Next, the atlas probabilities for 14 structures and the background are extracted, corresponding to all training samples and making a vector of size 15. These probabilities provide the network with spatial information and guide it to overcome intensity-based difficulties in some MRI volumes such as imaging artefacts and abnormalities caused by neurological diseases as black holes that appear next to the structures (Kushibar et al., 2018).

For the WMH lesion segmentation task, we used a cascaded training strategy as described in Valverde et al. (2017), where the network was trained in two stages. In the first step, the network is trained with a balanced number of samples extracted from all lesion voxels and an equal number of negative voxels randomly selected from non-lesion parts of the brain. Then, the same set of training images is segmented to obtain initial lesion masks. In the second stage, the network is also trained with a balanced

[2]https://wmh.isi.uu.nl.

[3]http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg.

[4]http://www.fmrib.ox.ac.uk/fsl.

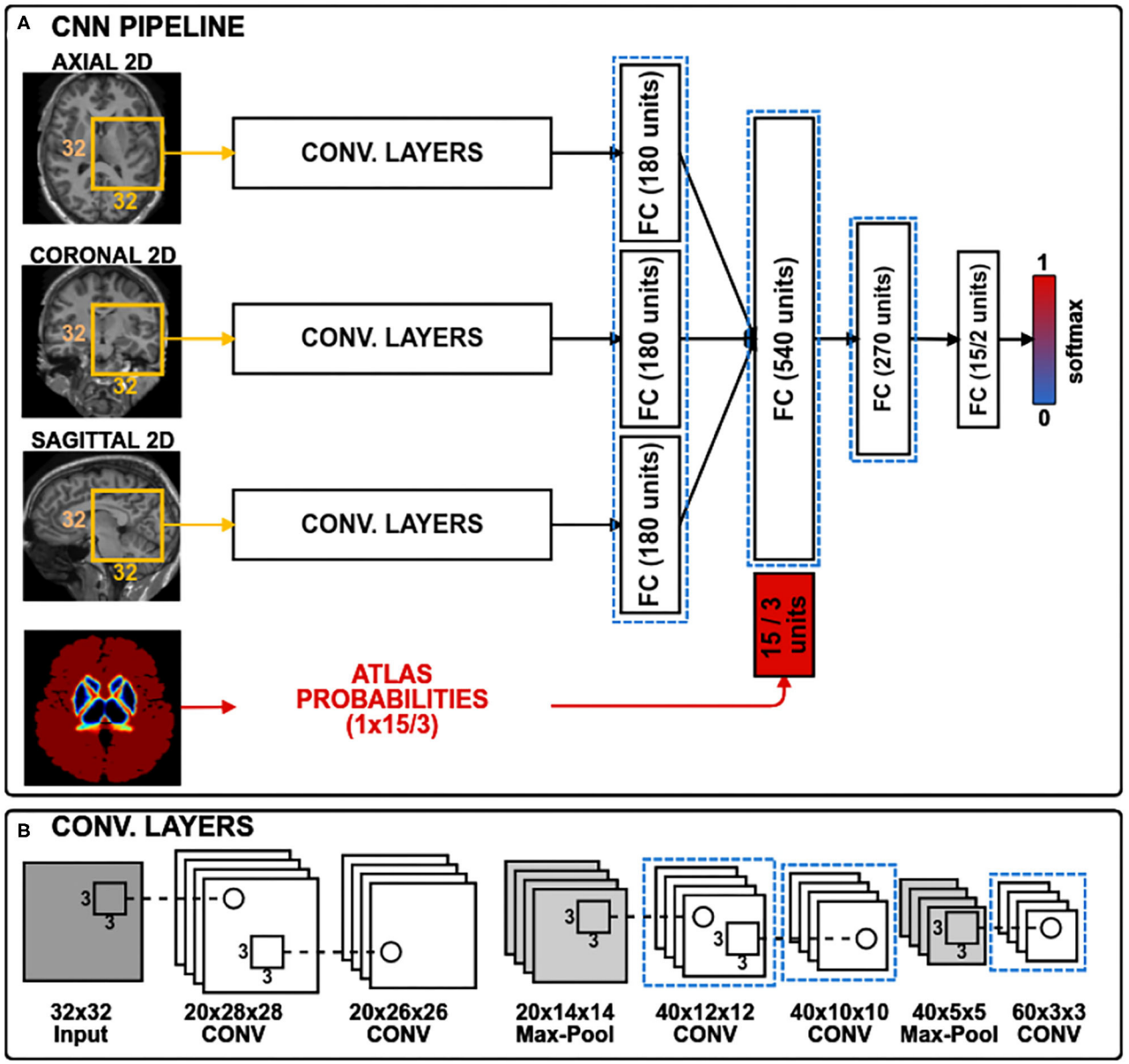


FIGURE 1 | The CNN architecture has three convolutional branches and a branch for spatial priors. 2D patches of size 32 × 32 pixels are extracted from three orthogonal views of a 3D volume. For sub-cortical structure segmentation, the spatial prior branch accepts a vector of size 15 with atlas probabilities for each of the 14 structures plus the background, whereas for the WMH lesion segmentation the vector size is three corresponding to white matter, gray matter and cerebrospinal fluid. Histogram loss is computed from the activation maps of the layers, highlighted with dashed blue rectangles. **(A)** CNN pipeline; **(B)** Convolutional layers.

set containing all lesion samples, however, the negative samples are extracted only from the voxels that were incorrectly classified in the first segmentation stage. This step is equivalent to a false positive reduction step.

For both tasks, the training samples were extracted along with their atlas probabilities, and randomly split into training and validation sets with 75 and 25% proportions, respectively. The training of the network was performed in batches of 128 for 200 epochs. An early-stopping protocol was defined with patience 20—i.e., the training stops if no increase was observed in the validation accuracy for 20 consecutive epochs. Optimisation was conducted for the categorical cross-entropy loss function using the Adam optimisation method (Kingma and Ba, 2014) with a learning rate of 10^{-2}.

3.4. Transductive Domain Adaptation

In the problem of domain adaptation we refer to source and target domains, where the former is the image domain with ground truth labels used in the initial training phase and the latter represents the new image domain without ground truth masks.

When looking at the activation maps of the convolutional layers extracted for source and target, we can observe the differences in intensity distributions as shown in **Figure 2**. As can be seen in **Figure 2A**, the magnitude of the activation maps for the source appear brighter compared to the target (**Figure 2C**). This demonstrates how the domain-shift problem affects the CNN in the feature level. Thus, the fully connected layers, which are used to mine these extracted features, cannot generalise to a different domain. When performing traditional transfer learning by re-training the last few layers of the network, we are adapting the fully connected part to better interpret the changes shown in **Figures 2A,C**. However, ground truth labels are not always available to perform such transfer learning for domain adaptation.

In this paper, we propose an alternative approach to traditional transfer learning by adapting the feature maps in the network instead of retraining the last few layers. **Figure 3** illustrates the transductive training process pipeline. First, features maps are extracted from several layers of the CNN for source and target training images. Then, the activation maps from the source domain are mapped to the features of target domain using a histogram matching technique. Next, we calculate the distance from the original source features to the histogram matched feature distributions. This difference is back-propagated as a histogram loss to encourage the network to produce feature maps similar to the target.

Let L_i be the layers of the CNN that we want to apply the histogram loss, and let us define A_i and B_i as the activation maps from the source and target samples for the *ith* layer, respectively. Then, the histogram loss is computed as:

$$\mathcal{L}_{hist} = \sum_{i}^{L} LogCosh(A_i, H(A_i, B_i)), \tag{1}$$

where, $H(\cdot, \cdot)$ is a function that applies a regular histogram mapping from source A_i to B_i target, and *LogCosh* is a logarithm of hyperbolic cosine that mostly works like the mean squared error but less affected by occasional large differences in the feature maps. In this form, the histogram loss is differentiable, and the loss can be computed easily by storing the histogram matched matrices for A_i in memory. Moreover, with this approach, the images from the target domain are included in training in a transductive manner in the feature level with no requirement for ground truth labels. An example of histogram matched feature maps of the source samples is shown in **Figure 2B**. Here, we can observe that the spatial integrity is the same as the original features (**Figure 2A**) and the intensity distribution is similar to the target features (**Figure 2C**).

Note that overall, we aim to minimise the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{hist}, \tag{2}$$

where $\mathcal{L}_{ce}$ is a cross-entropy loss and λ is a hyper-parameter to weight the effect of the histogram loss. The cross-entropy loss is computed using the source images with ground truth labels. Inclusion of this term is important to make the network learn to adapt to the changes in the feature maps after the histogram loss takes effect.

In our experiments, setting λ to be 1.0 showed the best results. Also, it has to be noted that the performance of the method was not very sensitive to the values within 1 ± 0.6. However, much larger or smaller values caused overshooting or diminished the effect of histogram loss during training. One could increase or decrease this weight out of the suggested range when applying for a different task that was not addressed in this study to change the influence of the histogram loss. The learning rate was reduced to 10^{-4} to avoid rapid weight updates. Applying histogram matching per sample could be limited due to the variance of histograms from different locations in the brain. Therefore, the histogram loss is computed over a batch—in our case batches of 32—hence, the loss is computed over a distribution rather than per sample, which we note as a necessary requirement. We empirically chose the last three convolutional, and all fully connected layers except for the last classification layer to compute the histogram loss as shown in **Figure 1** with dashed blue rectangles. For both segmentation tasks, using only one image from source and target sets was sufficient to perform the domain adaptation.

3.5. Network Testing

To perform a segmentation with a trained model, all 2.5D patches and corresponding atlas probabilities are extracted from an MRI volume, then passed through the CNN to obtain a probability map for each patch.

For the sub-cortical structure segmentation, the final label is defined using the *argmax* function. For this task, we used patches only from a region of interest (ROI) defined by a mask from the dilated atlas probabilities of the structures. In doing so, we were able to speed up the segmentation process drastically because the sub-cortical structures are located in the central part of the brain. Since the network is well trained to classify the borders of the structures, there may appear some wrongly classified voxels, which are removed by keeping only the largest volume for each class.

For WMH lesion segmentation, we use all the available brain patches because lesions can be in any place in the brain within the white matter. The obtained output probability maps from the CNN are thresholded to produce binary outputs with lesion candidates. Then, all lesion candidates that are outside the white matter defined by the registered probabilistic atlas, as well as candidates that have a volume less than 3 mm^3 are removed (Filippi et al., 2016).

3.6. Experiments and Evaluation

In this section, we describe the experimental setups used to test our approach for the two different segmentation tasks.

For the sub-cortical structure segmentation problem, we set up two pre-trained baseline models with MICCAI 2012 and IBSR dataset images as source. Then, domain adaptation was carried out in three ways: (1) from IBSR baseline to MICCAI 2012; (2) from MICCAI 2012 baseline to IBSR-GE; and (3) from MICCAI 2012 baseline to IBSR-SIEMENS. We separated the IBSR dataset into the IBSR-GE and IBSR-SIEMENS sub-groups according to

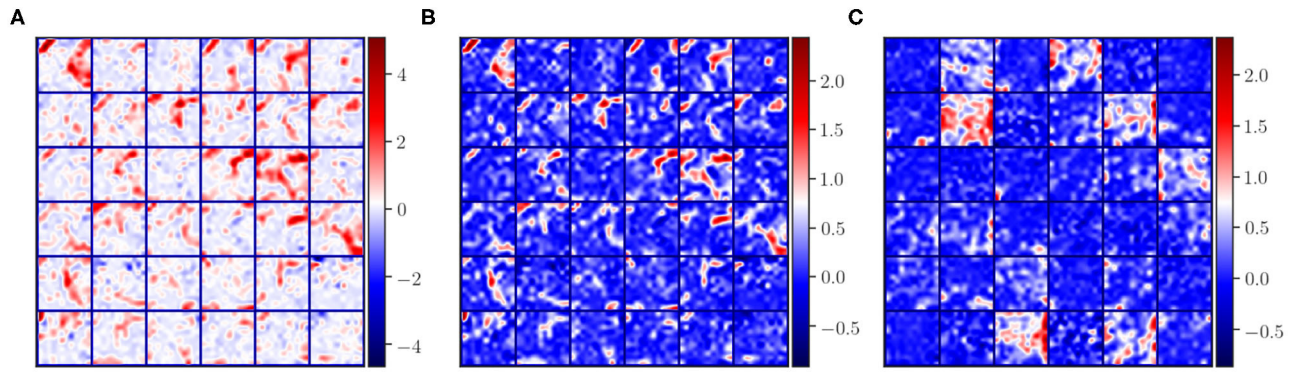


FIGURE 2 | Illustration of some activation maps for **(A)** source, **(B)** source after applying histogram matching to the target, and **(C)** target. Here, 36 example activation maps from the third convolutional layer are shown with the *"seismic"* color-map to visually emphasise the differences in magnitudes of the activation maps.

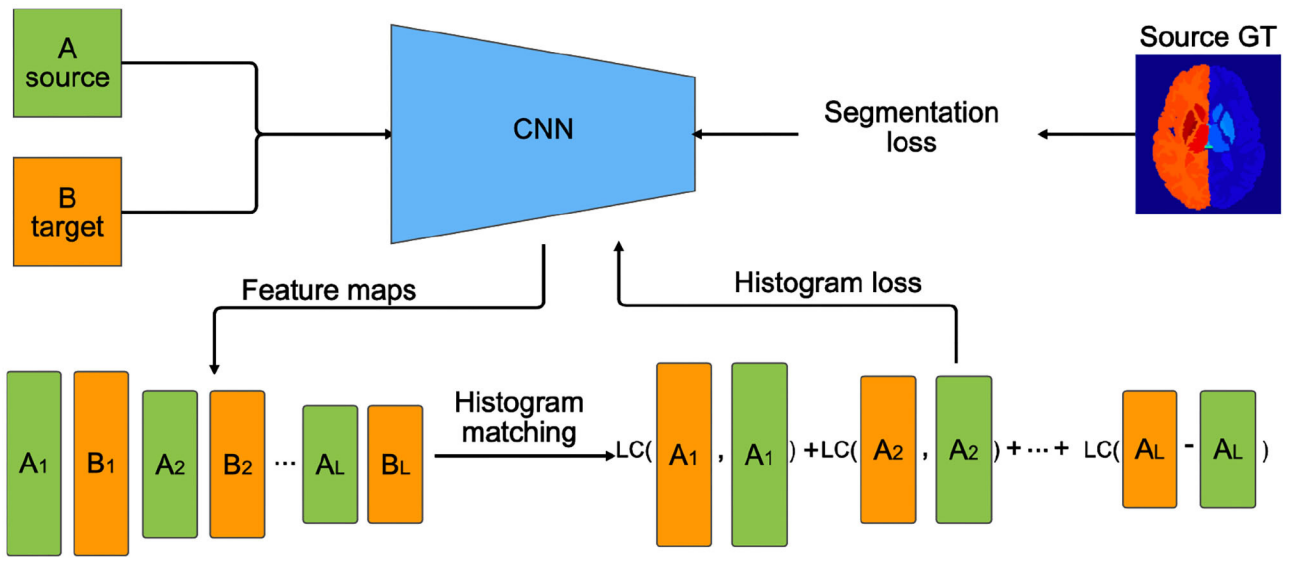


FIGURE 3 | Transductive domain adaptation training pipeline using histogram loss. A_i and B_i are feature maps extracted from the *ith* layers of the CNN. L is the number of layers on which the histogram loss is computed. Segmentation loss, in our case cross-entropy loss, is computed using the source ground truth (GT) labels. LC (LogCosh)—logarithm of hyperbolic cosine function.

the scanner manufacturer. This division was done to perform evaluation using the images with inter-scanner variability.

For the WMH lesion segmentation task we defined two pre-trained baseline models with WMH 2017 and VH dataset images as source. Then, we applied domain adaptation in four ways: (1) from WMH 2017 model to VH; (2) from VH model to UMC Utrecht site; (3) from VH model to Singapore site; and (4) from VH model to VU Amsterdam site.

Performing the domain adaptation for this experimental setup ensures that the source and target domains are different, and offers a realistic application of our proposal. We also compare our results with well-known unsupervised segmentation methods for both tasks. For the sub-cortical structure segmentation, we used the FSL-FIRST with default parameters, whereas for WMH lesion we used the LST method with κ thresholds empirically set to 0.4 and 0.1, which showed the best segmentation result for VH and WMH 2017 datasets, respectively.

For the sub-cortical structure segmentation task we reported the Dice Similarity Coefficient (DSC), since it is the most commonly used metric in the literature. The DSC is an overlap measurement that shows how well the automated segmentation is aligned with the gold standard; zero being no overlap and 1.0 full overlap. For the WMH lesion segmentation, along with the overlap DSC measure, we also used the common metrics of detection—True Positive Rate (TPR) and False Positive Rate (FPR)—which indicate the method's performance for detection and correct classification of the lesion candidates. Both the TPR and FPR values range between zero and one, where higher values are better for TPR and lower is better for FPR. Also, we used the common F-score metric that incorporates both measures to show

TABLE 1 | DSC results with standard deviations for the pre-trained baseline model without domain adaptation, transductive domain adaptation (TDA), and unsupervised FIRST method for two-way validation: from IBSR to MICCAI 2012; from MICCAI 2012 to IBSR-SIEMENS; and from MICCAI 2012 to IBSR-GE.

	IBSR to MICCAI 2012			**MICCAI2012 to IBSR-SIEMENS**			**MICCAI 2012 to IBSR-GE**		
	Baseline	**TDA**	**FIRST**	**Baseline**	**TDA**	**FIRST**	**Baseline**	**TDA**	**FIRST**
Tha.L	0.301 ± 0.195	0.843 ± 0.028*	**0.889 ± 0.017**	0.842 ± 0.029	0.873 ± 0.023	**0.892 ± 0.022**	0.681 ± 0.102	0.699 ± 0.111*	**0.894 ± 0.015**
Tha.R	0.085 ± 0.203	0.857 ± 0.022*	**0.890 ± 0.018**	0.823 ± 0.026	0.886 ± 0.016	**0.889 ± 0.014**	0.701 ± 0.108	0.736 ± 0.124*	**0.882 ± 0.011**
Cau.L	**0.867 ± 0.052**	0.861 ± 0.057	0.797 ± 0.117	0.862 ± 0.020	**0.887 ± 0.014**	0.805 ± 0.028	0.801 ± 0.074	**0.836 ± 0.046***	0.771 ± 0.047
Cau.R	**0.873 ± 0.040**	0.865 ± 0.044	0.837 ± 0.046	0.860 ± 0.011	0.864 ± 0.015	**0.892 ± 0.016**	0.828 ± 0.029	0.834 ± 0.025	**0.860 ± 0.026**
Put.L	0.888 ± 0.023	**0.893 ± 0.022**	0.860 ± 0.080	**0.891 ± 0.024**	0.888 ± 0.032	0.872 ± 0.016	0.852 ± 0.046	0.833 ± 0.053	**0.867 ± 0.023**
Put.R	0.887 ± 0.023	**0.889 ± 0.025**	0.876 ± 0.060	0.897 ± 0.008	**0.899 ± 0.013**	0.875 ± 0.011	0.842 ± 0.056	0.825 ± 0.064	**0.883 ± 0.009**
Pal.L	0.629 ± 0.083	0.785 ± 0.039*	**0.815 ± 0.060**	0.671 ± 0.048	0.737 ± 0.012	**0.827 ± 0.034**	0.557 ± 0.189	0.565 ± 0.182	**0.802 ± 0.031**
Pal.R	0.654 ± 0.058	0.768 ± 0.055*	**0.799 ± 0.088**	0.732 ± 0.053	0.785 ± 0.024	**0.808 ± 0.055**	0.574 ± 0.174	0.586 ± 0.175	**0.809 ± 0.028**
Hip.L	0.800 ± 0.025	**0.814 ± 0.029***	0.809 ± 0.014	0.804 ± 0.044	**0.813 ± 0.045**	0.811 ± 0.036	0.783 ± 0.037	0.797 ± 0.039	**0.804 ± 0.015**
Hip.R	0.832 ± 0.019	**0.839 ± 0.022***	0.810 ± 0.022	0.817 ± 0.049	**0.828 ± 0.053**	0.826 ± 0.034	0.795 ± 0.032	0.809 ± 0.031	**0.812 ± 0.014**
Amy.L	0.672 ± 0.041	0.685 ± 0.047	**0.721 ± 0.054**	0.630 ± 0.041	0.686 ± 0.053	**0.736 ± 0.090**	0.540 ± 0.130	0.601 ± 0.103*	**0.745 ± 0.050**
Amy.R	0.644 ± 0.056	0.671 ± 0.053*	**0.707 ± 0.052**	0.609 ± 0.074	0.637 ± 0.090	**0.756 ± 0.08**	0.455 ± 0.097	0.520 ± 0.088*	**0.758 ± 0.055**
Acc.L	0.695 ± 0.053	**0.707 ± 0.060**	0.699 ± 0.081	0.694 ± 0.050	**0.744 ± 0.036**	0.742 ± 0.069	0.646 ± 0.089	**0.658 ± 0.084**	0.655 ± 0.099
Acc.R	0.697 ± 0.067	**0.709 ± 0.070**	0.678 ± 0.089	0.634 ± 0.036	0.676 ± 0.042	**0.725 ± 0.063**	0.582 ± 0.081	0.595 ± 0.073	**0.691 ± 0.082**
Avg.	0.680 ± 0.038	**0.799 ± 0.087***	0.799 ± 0.094	0.769 ± 0.107	0.800 ± 0.094*	**0.818 ± 0.073**	0.688 ± 0.159	0.707 ± 0.147*	**0.802 ± 0.083**

Structure acronyms are: Tha.L, left thalamus; Tha.R, right thalamus; Cau.L, left caudate; Cau.R, right caudate; Put.L, left putamen; Put.R, right putamen; Pal.L, left pallidum; Pal.R, right pallidum; Hip.L, left hippocampus; Hip.R, right hippocampus; Amy.L, left amygdala; Amy.R, right amygdala; Acc.L, left accumbens; Acc.R, right accumbens; Avg., average value. Significant improvements after domain adaptation over baseline are indicated with "" and maximum DSC values are shown in bold.*

classifier accuracy in correctly detecting lesions, and it ranges from zero (low) to one (high).

We used the pairwise non-parametric Wilcoxon signed-rank test (two-sided) to compare the statistical significance of our results with respect to the results of the pre-trained baseline model without domain adaptation and the state-of-the-art tools. The results were considered significant for ($p < 0.05$). Moreover, we perform Bonferroni correction to the significance levels when comparing structure-wise and lesion-wise detection and segmentation for both of the selected tasks to counteract the multiple comparisons problem. Therefore, the differences will be assumed to be significant for ($p < 0.0036$) and ($p < 0.0125$) for sub-cortical structure and WMH lesion segmentation tasks, respectively.

All the experiments were run using a machine with a 3.40-GHz CPU clock and on a single TITAN-X GPU (NVIDIA corp, United States) with 12 GB of RAM memory. The network was implemented using the Keras (Chollet et al., 2018) deep learning library with Tensorflow backend[5].

4. RESULTS

4.1. Sub-cortical Structure Segmentation

Table 1 shows the DSC results of the pre-trained baseline model without domain adaptation, proposed domain adaptation method, and FIRST for three datasets. Also, **Figure 4** illustrates segmentation improvements from the baseline after applying domain adaptation with subject-wise correspondence of the volumes in the target dataset. When testing the method on the first set, where IBSR was source and MICCAI 2012 was target, significant improvement in the overall result was observed after applying the domain adaptation, reaching a DSC of 0.799 compared to the baseline segmentation with the DSC score of 0.680 ($p = 2.8 \times 10^{-27}$). The average DSC of our method was similar to FIRST and the difference was not statistically significant ($p = 0.160$). Significant structure-wise improvements in the results were also observed for most of the structures when domain adaptation was applied: left thalamus ($p = 8.9 \times 10^{-5}$), right thalamus ($p = 8.9 \times 10^{-5}$), left pallidum ($p = 8.9 \times 10^{-5}$), right pallidum ($p = 8.9 \times 10^{-5}$), left hippocampus ($p = 1.9 \times 10^{-4}$), right hippocampus ($p = 0.002$), and right amygdala ($p = 8.9 \times 10^{-5}$).

Significant improvement from 0.769 to 0.800 in overall DSC was achieved using the domain adaptation to the MICCAI 2012 baseline ($p = 2.1 \times 10^{-13}$), when tested on the IBSR-SIEMENS dataset. Also, improvements for most of the structures were observed compared to the baseline, however, not significant ($p > 0.0036$). The average DSC for FIRST was better compared to our method ($p = 0.008$), however, our domain adaptation method showed better or similar results for all structures, except for the pallidum and amygdala.

The second subset of the IBSR dataset (IBSR-GE) showed to be the most difficult to obtain better segmentation results as can be also seen in **Figure 4**, where the increase in DSC was smaller compared to other targets. However, significant improvements were achieved by using domain adaptation, improving the average DSC of the baseline from 0.688 to 0.707 ($p = 6.8 \times 10^{-10}$). Also, performance improvements were achieved for most of the structures and significant increases were observed for left thalamus ($p = 0.002$), right thalamus ($p = 0.0009$), left caudate

[5]https://www.tensorflow.org.

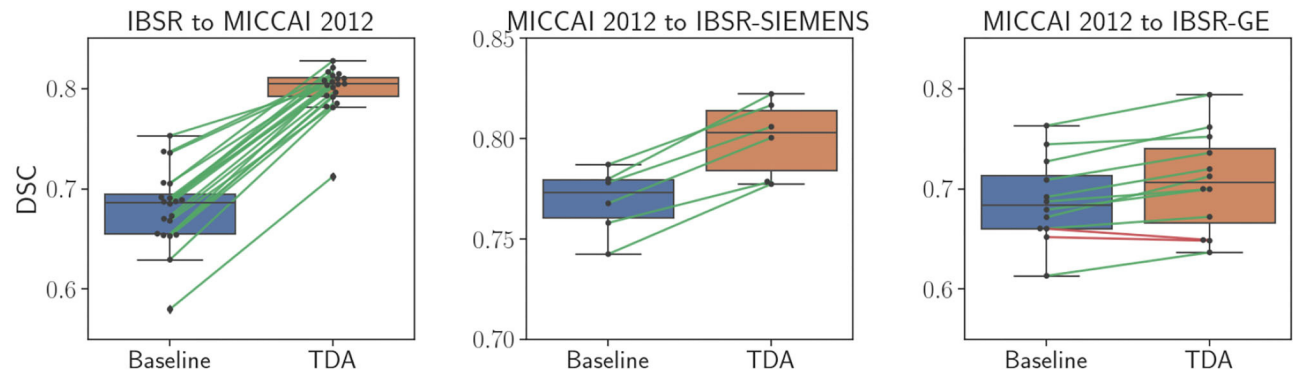


FIGURE 4 | Comparison of sub-cortical structure segmentation between direct testing (Baseline) and after domain adaptation (TDA). Black dots refer to each subject volume in the target dataset. The connecting lines show correspondence for improved (green) and decreased (red) DSC values.

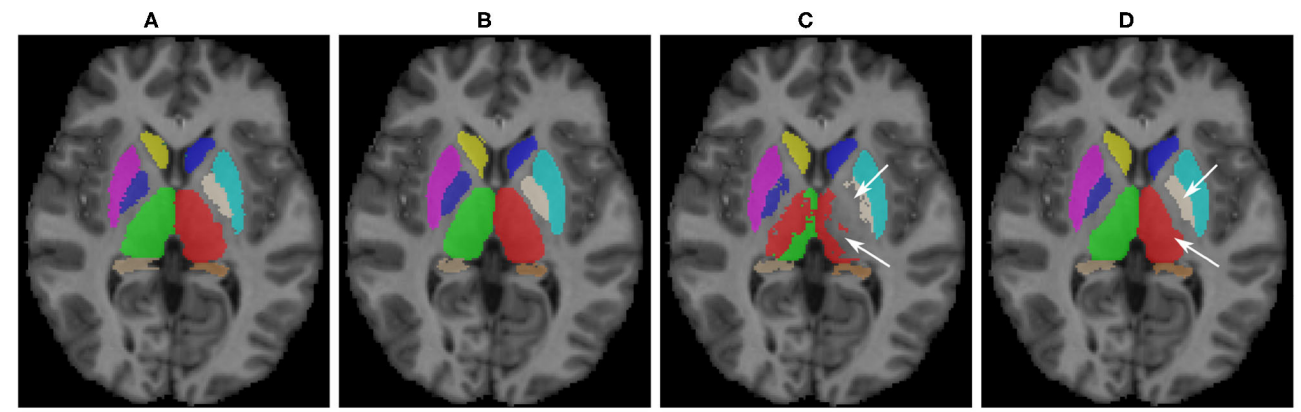


FIGURE 5 | Qualitative results for sub-cortical structure segmentation: **(A)** Ground truth; **(B)** FIRST segmentation; **(C)** Pre-trained baseline CNN output without domain adaptation; **(D)** After domain adaptation. Arrows indicate: top → pallidum; bottom → thalamus.

($p = 0.0005$), left amygdala ($p = 0.0009$), and right amygdala structures ($p = 0.0004$). The average DSC of FIRST (0.802) was significantly higher than our approach ($p = 1.7 \times 10^{-15}$) and similar behaviour was observed for most of the structures. Similar outcome with this sub-group of the IBSR dataset has also been noticed in Kushibar et al. (2019) which will be further discussed in section 5.

Some qualitative results are shown in **Figure 5** for the MICCAI 2012 dataset image as target. As can be seen, the baseline model did not produce satisfactory segmentation results for the thalamus and pallidum structures (indicated with arrows), which were improved after the domain adaptation. The proposed transductive domain adaptation method for segmentation greatly improved the model's performance and alleviated the segmentation errors caused by the domain-shift.

The training time for this task was 11 min on average per epoch. Additionally, the segmentation time using our method was 1.3 min (run on GPU) + 3.7 min (atlas registration, run on CPU) per volume on average. In contrast, FIRST took 10 min on average to segment all the sub-cortical structures in one subject volume.

We also tested the proposed method with the well-known U-Net architecture (Ronneberger et al., 2015) by applying the histogram loss in the features of the bottleneck layer. The average DSC for MICCAI 2012 dataset for baseline and after domain adaptation was 0.815 ± 0.097 and 0.816 ± 0.087, respectively. Similarly, the SIEMENS subset of the IBSR dataset yielded a DSC of 0.791 ± 0.103 and 0.790 ± 0.110 for baseline and TDA, respectively. A slight improvement was observed in DSC for the GE subset increasing the average from 0.738 ± 0.129 to 0.756 ± 0.115. A more detailed analysis will be discussed in section 5.

4.2. WMH Lesion Segmentation

Table 2 shows quantitative results for the WMH lesion segmentation using the pre-trained baseline model without domain adaptation, our proposed domain adaptation method, and the unsupervised method LST. Additionally, **Figure 6** illustrates segmentation improvements with subject-wise correspondence between baseline and domain adaptation methods for the subject volumes of the target dataset.

When the WMH 2017 dataset was used as source and VH as target, a significant improvement was achieved in segmentation,

TABLE 2 | WMH lesion segmentation results for the pre-trained baseline model without domain adaptation, transductive domain adaptation (TDA), and unsupervised LST method for four different sites: (1) source WMH 2017 and VH target; (2) source VH to Singapore; (3) source VH to UMC Utrecht; and (4) source VH to VU Amsterdam.

WMH 2017 to VH (3T Siemens TrioTim)				VH to Singapore (3T Siemens TrioTim)			
	Baseline	**TDA**	**LST**		**Baseline**	**TDA**	**LST**
DSC	0.478 ± 0.229	**0.536 ± 0.232***	0.410 ± 0.232	DSC	0.636 ± 0.176	**0.703 ± 0.198***	0.651 ± 0.176
TPR	0.735 ± 0.208	0.544 ± 0.231	0.319 ± 0.210	TPR	0.314 ± 0.089	0.451 ± 0.106	0.148 ± 0.092
FPR	0.611 ± 0.226	0.480 ± 0.256	0.477 ± 0.273	FPR	0.211 ± 0.186	0.469 ± 0.197	0.510 ± 0.153
F-score	0.270 ± 0.186	**0.308 ± 0.187***	0.160 ± 0.140	F-score	0.265 ± 0.102	**0.289 ± 0.118**	0.106 ± 0.067
VH to UMC Utrecht (3T Philips Achieva)				**VH to VU Amsterdam (3T GE Signa)**			
	Baseline	**TDA**	**LST**		**Baseline**	**TDA**	**LST**
DSC	0.587 ± 0.203	**0.624 ± 0.210***	0.620 ± 0.201	DSC	0.504 ± 0.148	**0.602 ± 0.135***	0.581 ± 0.155
TPR	0.464 ± 0.107	0.464 ± 0.148	0.250 ± 0.130	TPR	0.478 ± 0.114	0.483 ± 0.106	0.290 ± 0.105
FPR	0.279 ± 0.151	0.319 ± 0.175	0.352 ± 0.221	FPR	0.284 ± 0.155	0.298 ± 0.184	0.358 ± 0.161
F-score	0.316 ± 0.103	**0.318 ± 0.111**	0.181 ± 0.091	F-score	0.300 ± 0.108	**0.341 ± 0.126***	0.213 ± 0.095

DSC, dice similarity coefficient; TPR, true positive rate; FPR, false positive rate. Highest DSC and F-scores are shown in bold. Statistically significant improvements from baseline are indicated with "".*

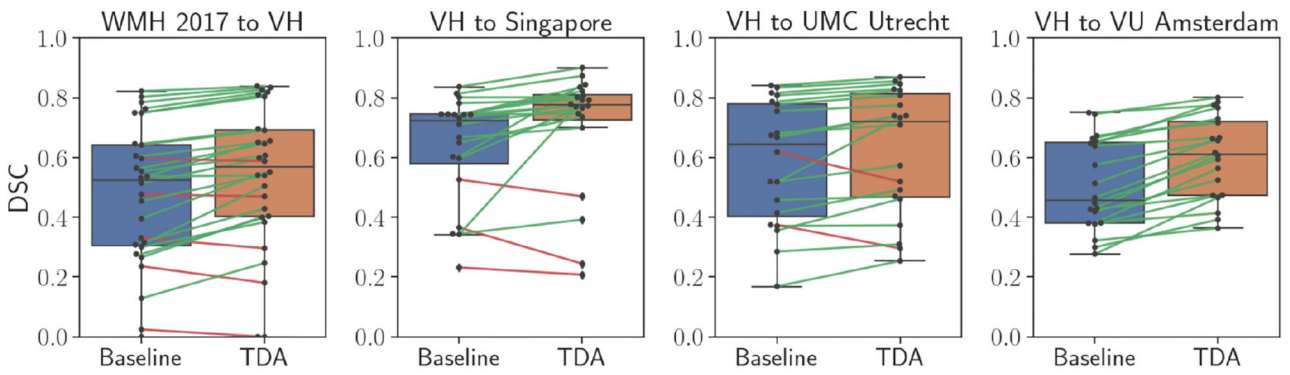


FIGURE 6 | Comparison of WMH lesion segmentation between direct testing (Baseline) and after domain adaptation (TDA). Black dots refer to each subject volume in the target dataset. The connecting lines show correspondence for improved (green) and decreased (red) DSC values.

increasing the DSC from 0.410 to 0.536 ($p = 0.0002$). The F-score was significantly improved from 0.270 to 0.308 ($p = 0.007$) as was the FPR, significantly improving from 0.611 to 0.480 ($p = 2.9 \times 10^{-5}$), however, there was a decrease in TPR from 0.735 to 0.544 due to inter-rater variability, which will be further discussed in detail (section 5). In comparison to the DSC result for LST (0.410) and the F-score of 0.160, our method yielded significantly higher DSC ($p = 0.001$) and detection rates ($p = 2.4 \times 10^{-5}$) at similar operating points.

Significant improvements were obtained in lesion segmentation after applying domain adaptation from the pre-trained baseline without domain adaptation to the Singapore site, increasing the DSC from 0.636 to 0.703 ($p = 0.006$). A slight improvement was achieved in F-score but not statistically significant ($p = 0.156$). In comparison to LST, our method was significantly better in both segmentation and detection, ($p = 0.006$) and ($p = 0.0002$), respectively.

Performing domain adaptation from source VH to the target UMC Utrecht site significantly improved the DSC from 0.587 of baseline to 0.624 ($p = 0.008$). There were no improvements in lesion detection rates, and the differences in F-scores for the baseline and domain adaptation were not statistically significant ($p = 0.794$). The DSC using our method was similar to that of LST (0.620), and differences were not significant ($p = 0.79$), but significantly higher lesion detection rate was observed after domain adaptation in comparison to LST ($p = 0.0003$).

When the VU Amsterdam site was used as target, our approach achieved a significant increase in DSC, improving the baseline from 0.504 to 0.602 ($p = 8.9 \times 10^{-5}$). The F-score of our method with 0.341 was also significantly higher than both LST ($p = 0.0006$) and baseline ($p = 0.0002$) values, with 0.213 and 0.300, respectively. The segmentation performance of our method was slightly better than LST but not statistically significant ($p = 0.433$).

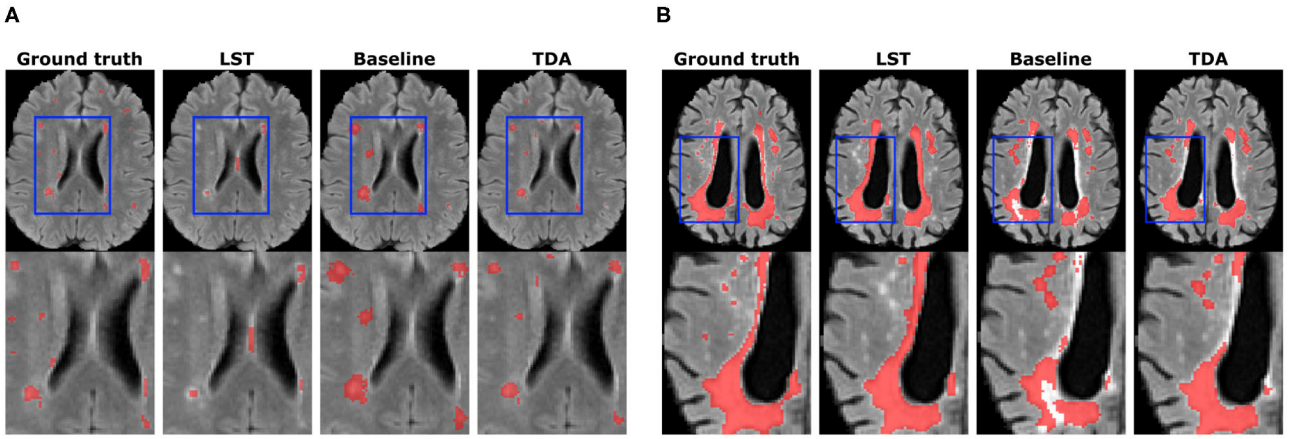


FIGURE 7 | Qualitative results for WMH lesion segmentation. Small lesion **(A)** and large lesion **(B)** segmentation improvements are shown. The bottom row depicts zoomed regions of interests shown in blue rectangles on whole-brain images (top row).

Figure 7 illustrates WMH lesion segmentation examples for the pre-trained baseline without domain adaptation, after transductive domain adaptation, and unsupervised LST. As can be seen, our method produced more refined segmentation than the baseline and better detection of smaller lesions. On the other hand, LST produced more false negatives and false positives for the smaller lesions. Some false negatives for the small lesions could not be avoided even after applying domain adaptation.

In comparison to the sub-cortical structure segmentation, the number of voxels in training was varying depending on the lesion load in the source image. Since the ground truth labels are available for the source images, we handpicked a representative image with a large lesion load. It took 14 min on average per training epoch. Furthermore, the segmentation time per volume using our method was 4 min (run on GPU) + 3 min (atlas registration, run on CPU) on average. Whereas LST took 25 min on average to segment the WMH lesions in one subject volume.

5. DISCUSSION

In this paper, we have introduced a novel domain adaptation method which minimises the differences in activation maps between the source and target domains in a transductive manner. As shown in **Figure 2**, the convolutional layers of the CNN produce different intensity distributions due to the variations in MRI images with different acquisition protocols. In order to alleviate this domain-shift effect, we performed histogram matching on the activation maps for the last convolutional layers as well as the fully connected layers of the network (**Figure 1**).

In the transductive domain adaptation process, we consider that manual annotations are only available for the source images, hence, optimisation of the CNN for segmentation loss can be only done using the source dataset. Therefore, the histograms of the activation maps extracted from the source were matched to those of the target. Then, the histogram loss function (Equation 1) computes how far the source feature map distributions are from the ones of target. In this way, the layers of the network are trained to produce similar activation maps to the target to minimise the distribution differences between two domains and jointly training the network to classify the input patches.

As can be seen in the results for the sub-cortical structure segmentation (**Table 1**), the performance of the pre-trained baseline CNN without domain adaptation was low. Moreover, this could also be observed in the segmentation example for one of the MICCAI 2012 dataset images (**Figure 5**), where the thalamus and pallidum structures were difficult for the network to segment. This is due to the weaker contrast between the structure boundaries and the background in comparison to other sub-cortical structures. On the other hand, the baseline segmentation for the putamen structure was better even for the baseline model. Although significant improvements were observed for both left and right putamen structures when using our domain adaptation method for the MICCAI 2012 dataset, this was not the case for the IBSR-SIEMENS and IBSR-GE datasets. However, the performance of the baseline model was similar to the one for transfer learning (Kushibar et al., 2019) due to the high contrast that this structure has compared to the background, which makes it easier for the network to generalise between different protocols. Aside from the putamen structure, our method was effective in improving the performance of the CNN for all other structures and significantly improved overall average DSC from the baseline.

The performance of the network after domain adaptation was similar to that of FIRST for the MICCAI 2012 dataset and slightly lower for the IBSR-SIEMENS dataset. However, for the IBSR-GE dataset, the result of domain adaptation was lower than that of FIRST. The MRI scans of IBSR-GE have imaging artefacts and lower quality in terms of contrast and brightness, which makes this subset of the IBSR dataset the most challenging one. In fact, the result of supervised domain adaptation using transfer learning with one image (0.784) was still lower than that of FIRST, and according to Kushibar et al. (2019), it took three images to significantly outperform FIRST using transfer learning. Since FIRST is an active-shape based model it is more

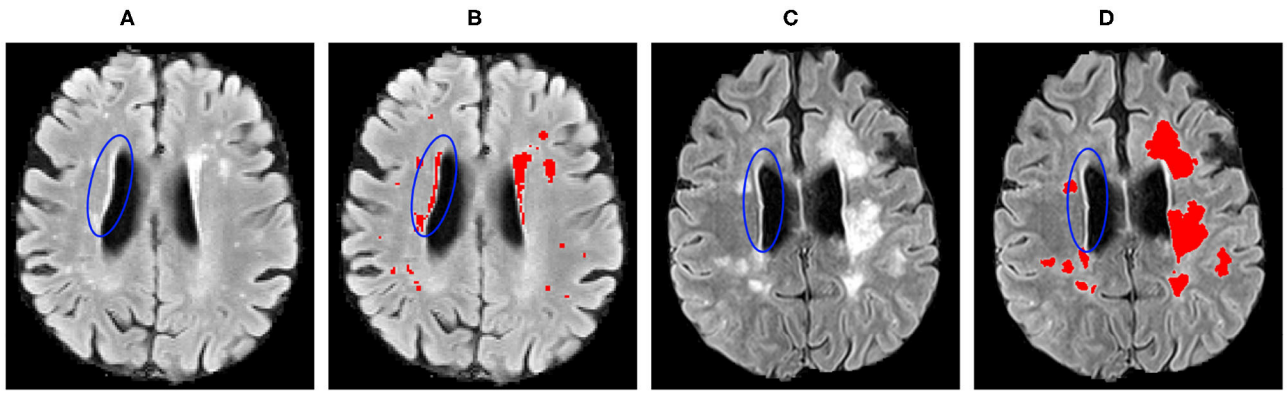


FIGURE 8 | Inter-operator variability in the lesion ground truth masks for the: **(A,B)** WMH 2017; and **(C,D)** VH datasets. Blue ellipses indicate the hyperintense tissues near the ventricles.

robust to imaging artefacts such as motion, and can produce moderate results despite the present difficulties. However, deep-learning based supervised methods (Dolz et al., 2018; Wachinger et al., 2018; Liu et al., 2020) outperform unsupervised ones if an adequate number of images are used in training.

The proposed method showed similar improvements when performing domain adaptation from pre-trained baseline model in the results for the WMH lesion segmentation task (**Table 2**). In general, significant improvements were observed in segmentation for all the experiments, while lesion detection was improved for some sites only. We have noticed that for this segmentation problem, inter-operator variability in the gold-standard lesion masks has an enormous effect on the lesion detection. As can be seen in **Figure 8**, the periventricular hyperintensities are annotated as lesions for the WMH 2017 dataset and not in VH. Moreover, there are more smaller lesions in the WMH 2017 dataset compared to the VH that have images with predominantly larger lesions. These differences introduce more difficulties in terms of better generalisation for the network and require supervised intervention to mitigate the problems of inter-operator differences between datasets.

Apart from these challenges, as shown in **Figure 7**, the proposed domain adaptation method significantly improved the segmentation result and produced better delineations of the lesion boundaries. Also, some smaller lesions were detected better after the domain adaptation, but some false positives still could not be avoided.

As shown in **Table 2**, adapting the network from WMH 2017 to the VH dataset significantly improved overall segmentation and detection rates. Also, for the images of the VH site, the results for both the baseline and domain adaptation were better than that of LST in terms of segmentation and lesion detection. However, for all the other target sites, we observed that the pre-trained baseline model without domain adaptation performed worse than LST and considerable improvements were achieved after applying domain adaptation. Overall, when adapting the model from VH to the different sites of the WMH 2017 datasets, lesion detection was not improved substantially. This was due to the inter-operator differences in the ground truths, where the CNN model was specifically trained to classify the small and periventricular hyperintense tissues as the background. However, as could be seen in **Figure 6**, segmentation performance was increasing for most of the subjects after applying the domain adaptation. We observed no improvement or decline in DSC for some subjects when the performance of the baseline was also low. Additionally, we observed that having at least the same scanner makes the network to be less affected by the domain shift. This could be seen in the example of NUHS Singapore site, which shares the same scanner as VH, but with different voxel resolution.

In terms of the number of images, our experiments showed that using only one image was enough for domain adaptation. This is because the histogram loss is computed only over the image features and the number of overall samples was adequate for the network to converge for both tasks. Including more training images did not improve the segmentation results due to the inter-operator variability in the expert annotated ground truths labels but increased the training time.

As could be seen in both the quantitative and qualitative results, the proposed transductive domain adaptation method is an effective way to mitigate the problems of domain-shift without the requirement for expert annotated labels. However, there are some limitations for domain adaptation when no ground truth labels are available. As we have seen in the results for the sub-cortical structure segmentation, transductive domain adaptation did not improve the DSC for structures where the performance of the pre-trained baseline model was already satisfactory to a certain degree. Similar behaviour was also observed when applying the proposed method with a commonly used U-Net architecture where the results were similar to the baseline for the MICCAI 2012 and IBSR-SIEMENS datasets. However, there was a slight improvement in the case of the IBSR-GE dataset where the baseline was affected by domain shift compared to the other sets. In general, we have noticed that U-Net was less affected by domain shift compared to our selected CNN. Moreover, it could be that the encoder-decoder architecture makes it difficult to perform TDA at the feature-level. However, the overall performance of U-Net when trained from scratch was lower

than that of the 2.5D approach that achieves the state-of-the-art results for the sub-cortical structure segmentation (Avg DSC 0.85 vs. 0.87, for UNet and our method in MICCAI 2012 dataset, respectively). Further investigation on improving the feature-level domain adaptation in encoder-decoder architectures with our proposed transductive method will be taken as a future work.

Furthermore, the inter-operator variability between two datasets also makes it challenging to evaluate such approaches. We recommend applying the transductive approach for domain adaptation to overcome extreme performance drops caused by domain-shift, and when there are no manually annotated images available. Although manually annotating the MRI scans for both considered segmentation problems is a time-consuming task, supervised transfer learning approaches remain a better way to address the domain-shift problem which could be better than the traditional unsupervised methods.

In general, most of the methods in the literature address domain adaptation where the source and target images are drastically different. Moreover, there are benchmark datasets that allow such comparisons in computer vision [for example, MNIST to The Street View House Numbers (SVHN)], but we still lack such standard datasets in the medical domain. We believe some medical benchmark datasets with minimal inter-operator variability in the ground-truths masks will emerge. For example, the iSeg infant brain tissue segmentation challenge (Sun et al., 2020) and the MnM Challenge for multi-site and multi-vendor cardiac MRI segmentation (Campello and Lekadir, 2020) have recently been organised addressing this challenge. Such initiatives would definitely serve as a benchmark for domain adaptation methods. Especially for the cases when the differences in images are not drastic but still affect the performance of deep learning based methods. Also, note that in Sun et al. (2020), the reported top five methods did not propose any domain adaptation method, and the ones utilising adversarial training or CycleGAN based approaches were not among the top methods, which shows how challenging the problem is. Although these more complex methods have shown their effectiveness in multi-modality setup, there is still room for improvement in domain adaptation for multi-site single-modality cases.

6. CONCLUSIONS

In this paper, we have introduced a transductive transfer learning method for reducing the domain-shift effect in deep learning caused by differences in MRI scanners and image-acquisition parameters. In our approach, we computed the histogram loss defined by the differences in the histogram distributions of the activation maps for the source and target domains from the convolutional and fully connected layers of the network. Minimising the histogram loss forces the convolutional layers to produce outputs for the source which are similar to those of the target. The network is end-to-end trainable and does not require exhaustive hyper-parameter tuning.

In order to implement our pipeline, we used a network architecture recently proposed in Kushibar et al. (2018), which had shown state-of-the-art performance in sub-cortical brain structure segmentation. We employed this architecture to perform domain adaptation for two different segmentation problems. The proposed approach was tested with different experimental setups using inter-site and inter-scanner datasets.

The experimental results confirmed the effectiveness of our domain adaptation approach for two different segmentation problems, where it was possible to significantly improve the performances of the pre-trained baseline models. Performing similarly to state-of-the-art traditional unsupervised methods, our approach was able to overcome extreme performance drops caused by domain-shift problem and achieve faster segmentation process. Moreover, along with the domain-shift issue, there are differences in the manual segmentation masks, which makes evaluation of domain adaptation pipelines more challenging.

In summary, the approach presented in this work, can help to improve brain biomarker extraction for various neurological and neurodegenerative disorders, especially in clinical scenarios where manual annotation are not available. Additionally, we have made our transductive transfer learning domain adaptation pipeline available to the research community at https://github.com/NIC-VICOROB/sub-cortical_segmentation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. This data can be found at: https://www.oasis-brains.org/#data; https://www.nitrc.org/projects/ibsr; https://wmh.isi.uu.nl.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Alex Rovira, Magnetic Resonance Unit, Department of Radiology, Vall d'Hebron University Hospital, Spain. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

KK: methodology, experiments, and writing. MS, SV, and JS: validation and review. ÀR: data provision and validation. AO and XL: supervision and review. All authors contributed to the article and approved the submitted version.

FUNDING

KK holds FI-DGR2017 grant from the Catalan Government with reference number 2017FI_B00372. This work has been supported by DPI2017-86696-R from the Ministerio de Ciencia y Tecnologia.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN-X PASCAL GPU used in this research.

REFERENCES

Ackaouy, A., Courty, N., Vallée, E., Commowick, O., Barillot, C., and Galassi, F. (2020). Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from MRI data. *Front. Comput. Neurosci.* 14:19. doi: 10.3389/fncom.2020.00019

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *J. Digit. Imaging* 30, 449–459. doi: 10.1007/s10278-017-9983-4

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [Preprint]. arXiv:1811.02629*.

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., et al. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artif. Intell. Med.* 95, 64–81. doi: 10.1016/j.artmed.2018.08.008

Campello, M., and Lekadir, K. (2020). "Multi-centre multi-vendor & multi-disease cardiac image segmentation challenge (M&Ms)," in *Medical Image Computing and Computer Assisted Intervention*. (Lima).

Caviness, V. S. Jr., Meyer, J., Makris, N., and Kennedy, D. N. (1996). MRI-based topographic parcellation of human neocortex: an anatomically specified method with estimate of reliability. *J. Cogn. Neurosci.* 8, 566–587. doi: 10.1162/jocn.1996.8.6.566

CC-BY (2020). *About The Creative Commons Licenses*. Available online at: http://creativecommons.org/about/licenses

Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P. A. (2020). Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Trans. Med. Imaging* 39, 2494–2505. doi: 10.1109/TMI.2020.2972701

Chollet, F. (2018). *Deep Learning With Python*, Vol. 361. New York, NY: Manning.

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, (Munich), 447–463. doi: 10.1007/978-3-030-01225-0_28

De Jong, L., Van der Hiele, K., Veer, I., Houwing, J., Westendorp, R., Bollen, E., et al. (2008). Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain* 131, 3277–3285. doi: 10.1093/brain/awn278

Debette, S., and Markus, H. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* 341:c3666. doi: 10.1136/bmj.c3666

Dolz, J., Desrosiers, C., and Ayed, I. B. (2018). 3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study. *NeuroImage* 170, 456–470. doi: 10.1016/j.neuroimage.2017.04.039

Filippi, M., Rocca, M. A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., et al. (2016). MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol.* 15, 292–303. doi: 10.1016/S1474-4422(15)00393-2

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N., et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* 132, 198–212. doi: 10.1016/j.neuroimage.2016.02.036

Frazier, J. A., Chiu, S., Breeze, J. L., Makris, N., Lange, N., Kennedy, D. N., et al. (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *Am. J. Psychiatry* 162, 1256–1265. doi: 10.1176/appi.ajp.162.7.1256

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med. Image Anal.* 17, 1–18. doi: 10.1016/j.media.2012.09.004

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., et al. (2017). "Transfer learning for domain adaptation in MRI: application in brain lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 516–524. doi: 10.1007/978-3-319-66179-7_59

González-Villá, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R., and Lladó, X. (2016). A review on brain structures segmentation in magnetic resonance imaging. *Artif. Intell. Med.* 73, 45–69. doi: 10.1016/j.artmed.2016.09.001

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems, Vol. 27*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Montreal, QC: Curran Associates, Inc.), 2672–2680.

Houtchens, M., Benedict, R., Killiany, R., Sharma, J., Jaisani, Z., Singh, B., et al. (2007). Thalamic atrophy and cognition in multiple sclerosis. *Neurology* 69, 1213–1223. doi: 10.1212/01.wnl.0000276992.17011.b5

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R. G., and Landman, B. A. (2018). "Adversarial synthesis learning enables segmentation without target modality ground truth," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, (Washington, DC), 1217–1220. doi: 10.1109/ISBI.2018.8363790

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imaging* 30, 1617–1634. doi: 10.1109/TMI.2011.2138152

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., et al. (2017). "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International Conference on Information Processing in Medical Imaging* (Boone, NC: Springer), 597–609. doi: 10.1007/978-3-319-59050-9_47

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. (2019). "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Long Beach, CA), 4893–4902. doi: 10.1109/CVPR.2019.00503

Kennedy, D. N., Haselgrove, C., Hodge, S. M., Rane, P. S., Makris, N., and Frazier, J. A. (2012). CANDIShare: a resource for pediatric neuroimaging data. *Neuroinformatics* 10, 319–322. doi: 10.1007/s12021-011-9133-y

Kikinis, R., Shenton, M. E., Iosifescu, D. V., McCarley, R. W., Saiviroonporn, P., Hokama, H. H., et al. (1996). A digital brain atlas for surgical planning, model-driven segmentation, and teaching. *IEEE Trans. Visual. Comput. Graph.* 2, 232–241. doi: 10.1109/2945.537306

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv e-prints*.

Kuijf, H. J., Biesbroek, J. M., de Bresser, J., Heinen, R., Andermatt, S., Bento, M., et al. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities; results of the WMH segmentation challenge. *IEEE Trans. Med. Imaging.* 38, 2556–2568. doi: 10.1109/TMI.2019.2905770

Kushibar, K., Valverde, S., González-Villá, S., Bernal, J., Cabezas, M., Oliver, A., et al. (2018). Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Med. Image Anal.* 48, 177–186. doi: 10.1016/j.media.2018.06.006

Kushibar, K., Valverde, S., González-Villá, S., Bernal, J., Cabezas, M., Oliver, A., et al. (2019). Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Sci. Rep.* 9:6742. doi: 10.1038/s41598-019-43299-z

Kutzelnigg, A., Lucchinetti, C. F., Stadelmann, C., Brück, W., Rauschka, H., Bergmann, M., et al. (2005). Cortical demyelination and diffuse white matter injury in multiple sclerosis. *Brain* 128, 2705–2712. doi: 10.1093/brain/awh641

Landman, B., and Warfield, S. (2012). "MICCAI 2012 workshop on multi-atlas labeling," in *Medical Image Computing and Computer Assisted Intervention Conference*, (Nice).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, H., Loehr, T., Sekuboyina, A., Zhang, J., Wiestler, B., and Menze, B. (2020). Domain adaptive medical image segmentation via adversarial learning of disease-specific spatial patterns. *arXiv e-prints: arXiv-2001*.

Liu, L., Hu, X., Zhu, L., Fu, C.-W., Qin, J., and Heng, P.-A. (2020). ψ-Net: stacking densely convolutional LSTMs for sub-cortical brain structure segmentation. *IEEE Trans. Med. Imaging* 39, 2806–2817. doi: 10.1109/TMI.2020.2975642

Mak, E., Bergsland, N., Dwyer, M., Zivadinov, R., and Kandiah, N. (2014). Subcortical atrophy is associated with cognitive impairment in mild Parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis. *Am. J. Neuroradiol.* 35, 2257–2264. doi: 10.3174/ajnr.A4055

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., et al. (2010). Fast free-form deformation using graphics processing units. *Comput. Methods Prog. Biomed.* 98, 278–284. doi: 10.1016/j.cmpb.2009.09.002

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* 19, 143–150. doi: 10.1109/42.836373

Orbes-Arteainst, M., Cardoso, J., Sørensen, L., Igel, C., Ourselin, S., Modat, M., et al. (2019). "Knowledge distillation for semi-supervised domain adaptation," in *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, eds L. Zhou, D. Sarikaya, S. M. Kia, S. Speidel, A. Malpani, D. Hashimoto, M. Habes, T. Löfstedt, K. Ritter, H. Wang (Shenzhen: Springer), 68–76. doi: 10.1007/978-3-030-32695-1_8

Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* 56, 907–922. doi: 10.1016/j.neuroimage.2011.02.046

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann. Neurol.* 69, 292–302. doi: 10.1002/ana.22366

Rimol, L. M., Hartberg, C. B., Nesvåg, R., Fennema-Notestine, C., Hagler, D. J. Jr, Pung, C. J., et al. (2010). Cortical thickness and subcortical volumes in schizophrenia and bipolar disorder. *Biol. Psychiatry* 68, 41–50. doi: 10.1016/j.biopsych.2010.03.036

Rohlfing, T. (2012). Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging* 31, 153–163. doi: 10.1109/TMI.2011.2163944

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. (2017). "Stabilizing training of generative adversarial networks through regularization," in *Advances in Neural Information Processing Systems*, (Long Beach, CA), 2018–2028.

Rozantsev, A., Salzmann, M., and Fua, P. (2018). Beyond sharing weights for deep domain adaptation. *IEEE Trans. Pattern Anal. Mach. intell.* 41, 801–814. doi: 10.1109/TPAMI.2018.2814042

Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., et al. (2019). Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage* 23:101849. doi: 10.1016/j.nicl.2019.101849

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., et al. (2011). Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Med. Image Anal.* 15, 267–282. doi: 10.1016/j.media.2010.12.003

Storelli, L., Rocca, M. A., Pagani, E., Van Hecke, W., Horsfield, M. A., De Stefano, N., et al. (2018). Measurement of whole-brain and gray matter atrophy in multiple sclerosis: assessment with MR imaging. *Radiology* 2018:172468. doi: 10.1148/radiol.2018172468

Sun, Y., Gao, K., Wu, Z., Lei, Z., Wei, Y., Ma, J., et al. (2020). Multi-site infant brain segmentation algorithms: the iSeg-2019 Challenge. *arXiv [Preprint]. arXiv:2007.02096*. doi: 10.1109/TMI.2021.3055428

Valverde, S., Cabezas, M., Roura, E., González-Villá, S., Pareto, D., Vilanova, J. C., et al. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage* 155, 159–168. doi: 10.1016/j.neuroimage.2017.04.034

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentá, L., et al. (2019). One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *Neuroimage* 21:101638. doi: 10.1016/j.nicl.2018.101638

Van Norden, A. G., de Laat, K. F., Gons, R. A., van Uden, I. W., van Dijk, E. J., van Oudheusden, L. J., et al. (2011). Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. *BMC Neurol.* 11:29. doi: 10.1186/1471-2377-11-29

Wachinger, C., Reuter, M., and Klein, T. (2018). Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage* 170, 434–445. doi: 10.1016/j.neuroimage.2017.02.035

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 2019:101552. doi: 10.1016/j.media.2019.101552

Zhang, Y., Chen, H., Wei, Y., Zhao, P., Cao, J., Fan, X., et al. (2019). "From whole slide imaging to microscopy: deep microscopy adaptation network for histopathology cancer image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen: Springer), 360–368. doi: 10.1007/978-3-030-32239-7_40

Zhang, Y., Miao, S., Mansi, T., and Liao, R. (2018). "Task driven generative modeling for unsupervised domain adaptation: application to X-ray image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Granada: Springer), 599–607. doi: 10.1007/978-3-030-00934-2_67

Conflict of Interest: ÀR serves on scientific advisory boards for Novartis, Sanofi-Genzyme, Icometrix, SyntheticMR, and OLEA Medical, and has received speaker honoraria from Bayer, Sanofi-Genzyme, Bracco, Merck-Serono, Teva Pharmaceutical Industries Ltd, Novartis, Roche, and Biogen Idec.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 25 May 2021
doi: 10.3389/fnins.2021.662005

Harmonized Segmentation of Neonatal Brain MRI

Irina Grigorescu[1,2]*, *Lucy Vanes*[1,3], *Alena Uus*[1,2], *Dafnis Batalle*[1,4], *Lucilio Cordero-Grande*[1,2,5], *Chiara Nosarti*[1,3], *A. David Edwards*[1], *Joseph V. Hajnal*[1,2], *Marc Modat*[2] *and Maria Deprez*[1,2]

[1] Centre for the Developing Brain, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [2] Biomedical Engineering Department, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom, [3] Department of Child and Adolescent Psychiatry, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [4] Department of Forensic and Neurodevelopmental Science, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [5] Biomedical Image Technologies, ETSI Telecomunicación, Universidad Politécnica de Madrid & CIBER-BNN, Madrid, Spain

Deep learning based medical image segmentation has shown great potential in becoming a key part of the clinical analysis pipeline. However, many of these models rely on the assumption that the train and test data come from the same distribution. This means that such methods cannot guarantee high quality predictions when the source and target domains are dissimilar due to different acquisition protocols, or biases in patient cohorts. Recently, unsupervised domain adaptation techniques have shown great potential in alleviating this problem by minimizing the shift between the source and target distributions, without requiring the use of labeled data in the target domain. In this work, we aim to predict tissue segmentation maps on T_2-weighted magnetic resonance imaging data of an unseen preterm-born neonatal population, which has both different acquisition parameters and population bias when compared to our training data. We achieve this by investigating two unsupervised domain adaptation techniques with the objective of finding the best solution for our problem. We compare the two methods with a baseline fully-supervised segmentation network and report our results in terms of Dice scores obtained on our source test dataset. Moreover, we analyse tissue volumes and cortical thickness measures of the harmonized data on a subset of the population matched for gestational age at birth and postmenstrual age at scan. Finally, we demonstrate the applicability of the harmonized cortical gray matter maps with an analysis comparing term and preterm-born neonates and a proof-of-principle investigation of the association between cortical thickness and a language outcome measure.

Keywords: deep learning, segmentation, neonatal brain, unsupervised domain adaptation, cortical thickness

OPEN ACCESS

Edited by:
Diana M. Sima,
Icometrix, Belgium

Reviewed by:
Ashok Panigrahy,
University of Pittsburgh, United States
Adil Bashir,
Auburn University, United States

***Correspondence:**
Irina Grigorescu
irina.grigorescu@kcl.ac.uk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *31 January 2021*
Accepted: *21 April 2021*
Published: *25 May 2021*

Citation:
Grigorescu I, Vanes L, Uus A, Batalle D, Cordero-Grande L, Nosarti C, Edwards AD, Hajnal JV, Modat M and Deprez M (2021) Harmonized Segmentation of Neonatal Brain MRI. Front. Neurosci. 15:662005. doi: 10.3389/fnins.2021.662005

1. INTRODUCTION

Medical image deep learning has made incredible advances in solving a wide range of scientific problems, including tissue segmentation or image classification (Miotto et al., 2018). However, one major drawback of these methods is their applicability in a clinical setting, as many models rely on the assumption that the source and target domains are drawn from the same distribution. As a result, the efficiency of these models may drop drastically when applied to images which were

acquired with acquisition protocols different than the ones used to train the models (Kamnitsas et al., 2017; Orbes-Arteaga et al., 2019).

At the same time, combining imaging data from multiple studies and sites is necessary to increase the sample size and thereby the statistical power of neuroimaging studies. However, one major challenge is the lack of standardization in image acquisition protocols, scanner hardware, and software. Inter-scanner variability has been demonstrated to affect measurements obtained for downstream analysis such as voxel-based morphometry (Takao et al., 2011), and lesion volumes (Shinohara et al., 2017). Therefore, the purpose of harmonizing magnetic resonance imaging (MRI) datasets is to make sure that the differences arising from different image acquisition protocols do not affect the analysis performed on the combined data. For example, volumetric and cortical thickness measures should only be affected by brain anatomy and not the acquisition protocol or scanners.

A class of deep learning methods called domain adaptation (DA) techniques aims to address this issue by suppressing the domain shift between the training and test distributions. In general, DA approaches are either semi-supervised, which assume the existence of labels in the target dataset, or unsupervised, which assume the target dataset has no labels. For example, a common approach is to train a model on source domain images and fine-tune it on target domain data (Ghafoorian et al., 2017; Kushibar et al., 2019). Although these methods can give good results, they can become impractical as more often than not the existence of labels in the target dataset is limited or of poor quality. Unsupervised domain adaptation techniques (Ganin and Lempitsky, 2015; Kerfoot et al., 2019) offer a solution to this problem by minimizing the disparity between a source and a target domain, without requiring the use of labeled data in the target domain.

In our previous work (Grigorescu et al., 2020), we investigated two unsupervised DA methods with the aim of predicting brain tissue segmentations on 2D axial slices of T_2-weighted (T_2w) MRI data of an unseen neonatal population. We proposed an additional loss term in one of the methods, in order to constrain the network to more realistic reconstructions. Our models were trained using as source domain a dataset with majority of term-born neonates and as target domain a preterm-only population acquired with a different protocol. We calculated mean cortical thickness measures for every subject in the two datasets and we performed an ANCOVA analysis in order to find group differences between the predicted source and target domains. This analysis showed that our proposed method achieved harmonization of our two datasets in terms of cortical gray matter tissue segmentation maps. In this paper, we build on the aforementioned framework, which we expanded in three main ways. First, we build and train 3D neural networks in order to capture more information about the neonatal brain. Second, we extend the validation of our trained models to subsets of the two cohorts matched for gestational age (GA) at birth and postmenstrual age (PMA) at scan, for which we analyse tissue volumes and global and local cortical thickness (CT) measures. Finally, we perform an analysis comparing term and preterm-born neonates on the harmonized cortical gray matter maps and we show the importance of harmonizing the data by a proof-of-principle investigation of the association between cortical thickness and a language outcome measure.

2. MATERIALS AND METHODS

2.1. Data Acquisition and Preprocessing

The T_2w MRI data used in this study was collected as part of two independent projects: the developing Human Connectome Project (dHCP[1], approved by the National Research Ethics Committee REC: 14/Lo/1169), and the Evaluation of Preterm Imaging (ePrime[2], REC: 09/H0707/98) study. The dHCP neonates were scanned during natural unsedated sleep at the Evelina London Children's Hospital between 2015 and 2019. The ePrime neonates were scanned at the neonatal intensive care unit in Hammersmith Hospital between 2010 and 2013 (Edwards et al., 2018). Infants with major congenital malformations were excluded from both cohorts.

The dHCP data was acquired using a Philips Achieva 3T scanner and a 32-channels neonatal head coil (Hughes et al., 2017), using a T_2w turbo spin echo (TSE) sequence with fat suppression, and using the following parameters: repetition time $T_R = 12$ s, echo time $T_E = 156$ ms, TSE factor 12, and SENSE factors of 2.11 for the axial plane and 2.58 for the sagittal plane. Images were acquired with an in-plane resolution of 0.8×0.8 mm, slice thickness of 1.6 mm and overlap of 0.8 mm. For each volume, there was an acquisition of 125 slices in the transverse plane and 134 slices in the saggital plane. All data was motion corrected (Kuklisova-Murgasova et al., 2012; Cordero-Grande et al., 2018) and super-resolution reconstructed to a 0.5 mm isotropic resolution (Makropoulos et al., 2018).

The ePrime dataset was acquired with a Philips Intera 3T system and an 8-channel phased array head coil, using a T_2w TSE sequence with parameters: repetition time $T_R = 8.67$ s, echo time $T_E = 160$ ms, and TSE factor 16. Images were acquired with an in-plane resolution of 0.86×0.86 mm, slice thickness of 2 mm and overlap of 1 mm. For each volume, the acquisition ranged between 92 and 106 slices in the transverse plane.

Our two datasets comprise of 403 MRI scans of infants (184 females and 219 males) born between 23 and 42 weeks GA at birth and scanned at term-equivalent age (after 37 weeks PMA) as part of the dHCP pipeline, and a dataset of 486 MRI scans of infants (245 females and 241 males) born between 23 and 33 weeks GA and scanned at term-equivalent age as part of the ePrime project. **Figure 1** shows their age distribution.

Both datasets were pre-processed prior to being used by the deep learning algorithms. The ePrime volumes were linearly upsampled to 0.5 mm isotropic resolution to match the resolution of our source (dHCP) dataset. Both dHCP and ePrime datasets were rigidly aligned to a common 40 weeks gestational age atlas space (Schuh et al., 2018) using the MIRTK (Rueckert et al., 1999) software toolbox. Then, skull-stripping was performed on all of our data using the brain masks obtained with the Draw-EM pipeline for automatic brain MRI segmentation of the developing

[1] http://www.developingconnectome.org/

[2] https://www.npeu.ox.ac.uk/prumhc/eprime-mr-imaging-177

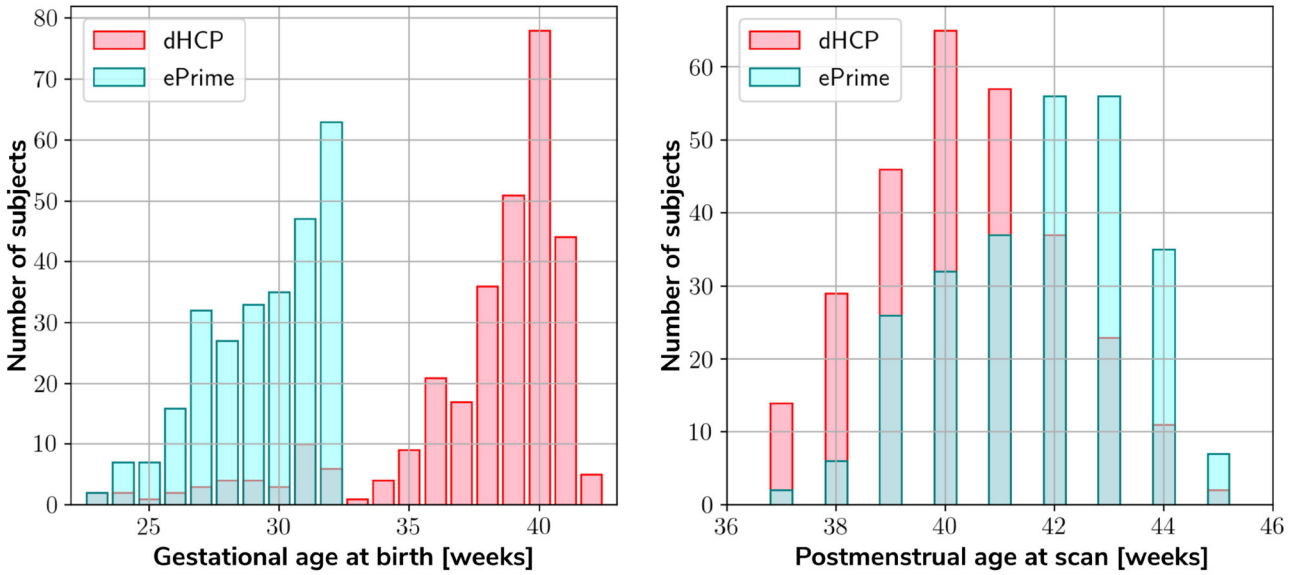


FIGURE 1 | Age distribution of the subjects in our datasets, showing both their GA at birth, as well as their PMA at scan.

neonatal brain (Makropoulos et al., 2018). Tissue segmentation maps were obtained using the same pipeline (Draw-EM) for both (dHCP and ePrime) cohorts.

To train our networks, we split our datasets into 80% training, 10% validation, and 10% test (see **Table 1**), keeping both the distribution of ages at scan and the male-to-female ratio as close to the original as possible. We used the validation sets to keep track of our models' performance during training, and the test sets to report our final models' results and showcase their capability to generalize.

TABLE 1 | Number of scans in different datasets used for training, validation and testing the models, together with their mean GA and PMA.

Dataset	#Subjects	GA at birth [weeks]	PMA at scan [weeks]
Train dHCP	340 (160♀ + 180♂)	39.1 (±2.7)	40.7 (±1.7)
Validate dHCP	32 (12♀ + 20♂)	39.3 (±1.6)	40.7 (±1.8)
Test dHCP	30 (12♀ + 19♂)	30 (±2.4)	41.4 (±1.7)
Train ePrime	417 (214♀ + 203♂)	29.6 (±2.3)	42.9 (±2.6)
Validate ePrime	38 (18♀ + 20♂)	29.8 (±2.3)	43 (±2.6)
Test ePrime	30 (13♀ + 18♂)	30 (±2.4)	41.4 (±1.7)

2.2. Unsupervised Domain Adaptation Models

To investigate the best solution for segmenting our target dataset (ePrime), we compared three independently trained deep learning models:

- **Baseline.** A 3D U-Net (Çiçek et al., 2016) trained on the source dataset (dHCP) only and used as a baseline segmentation network (see **Figure 2**).
- **Adversarial domain adaptation in the latent space.** A 3D U-Net segmentation network trained on source (dHCP) volumes, coupled with a discriminator trained on both source (dHCP) and target (ePrime) datasets (see **Figure 3**). This solution is similar to the one proposed by Kamnitsas et al. (2017) where the aim was to train the segmentation network such that it becomes agnostic to the data domain.
- **Adversarial domain adaptation in the image space.** Two 3D U-Nets, one acting as a generator, and a second one acting as a segmentation network, coupled with a discriminator trained on both real and synthesized ePrime volumes. The segmentation network is trained to produce tissue maps of the synthesized ePrime volumes created by the generator (see **Figure 4**). The normalized cross correlation (NCC) loss is added to the generator network to enforce image similarity between real and synthesized images, a solution which was previously proposed by Grigorescu et al. (2020).

To further validate the harmonized tissue maps, we trained an additional network (a 3D U-Net) to segment binary cortical tissue maps into 11 cortical substructures (see **Table 2**) based on anatomical groupings of cortical regions derived from the Draw-EM pipeline. The key reasons for training an extra network are: first, we avoid the time consuming task of label propagation between our available dHCP Draw-EM output segmentations and predicted ePrime maps, and second, we can train this network using Draw-EM cortical segmentations, and apply it on any brain cortical gray matter maps as in this case there will be no intensity shift between target and source distributions.

2.3. Network Architectures

The segmentation networks in all three setups and the generator used in the adversarial domain adaptation in the image space model have the same architecture, consisting of 5 encoding-decoding branches with 16, 32, 64, 128, and 256 channels, respectively. The encoder blocks use 3^3 convolutions (with

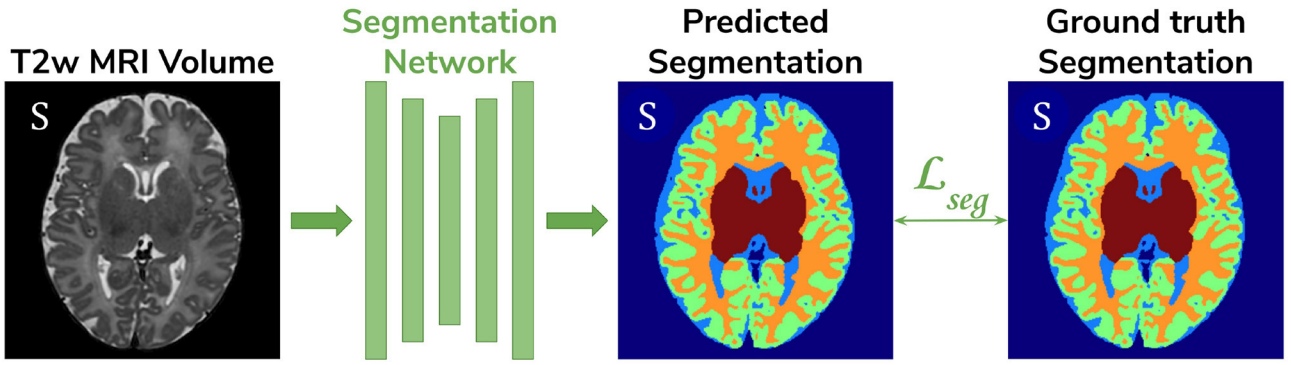


FIGURE 2 | The baseline model consists of a 3D U-Net trained to segment source (dHCP) volumes. The input T_2w MRI images, the predicted segmentation and the Draw-EM output segmentations are marked with S as they all belong to the source (dHCP) dataset.

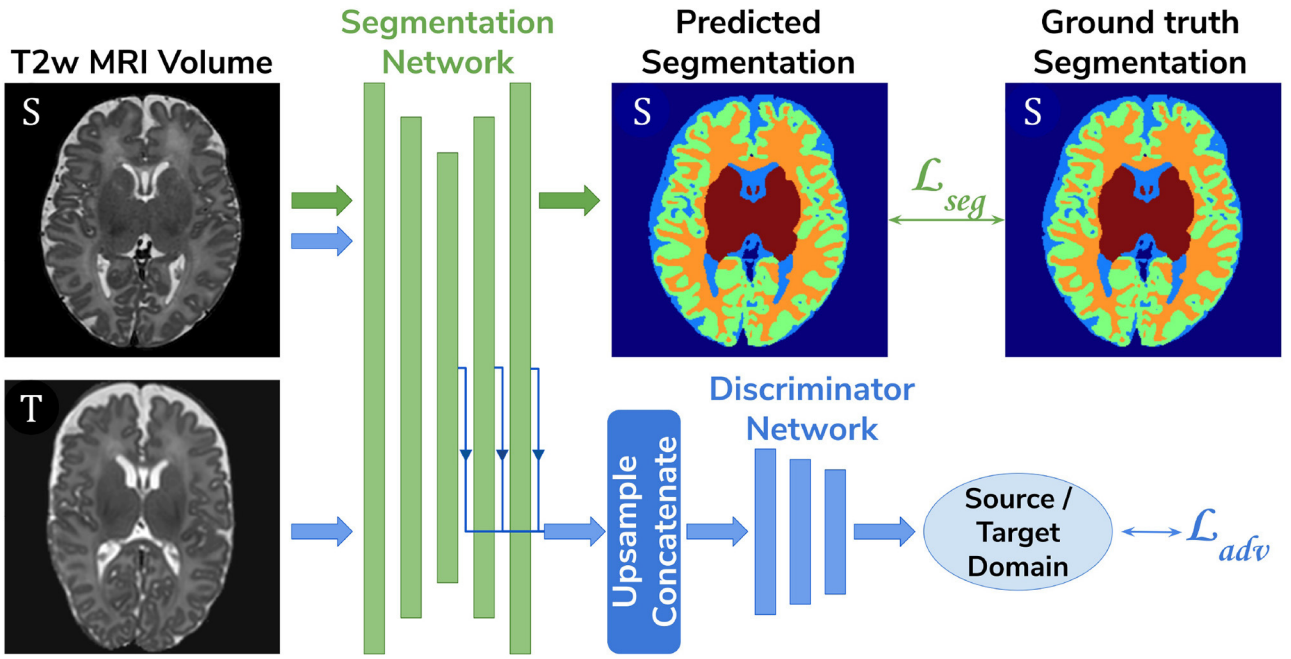


FIGURE 3 | The latent space domain adaptation setup consists of a 3D U-Net trained to segment the source (dHCP) T_2w MRI volumes, coupled with a discriminator network which forces the segmentation network to learn domain-invariant features. Both source (dHCP) and target (ePrime) images are fed to the segmentation network, but only source (dHCP) Draw-EM output labels are used to compute the segmentation loss. Source domain images are marked with S, while target domain images are marked with T, respectively.

a stride of 1), instance normalization (Ulyanov et al., 2016) and LeakyReLU activations. A 2^3 average pooling layer is used after the first down-sampling block, while the others use 2^3 max pooling layers. The decoder blocks consist of 3^3 convolutions (with a stride of 1), instance normalization (Ulyanov et al., 2016), LeakyReLU activations, and, additionally, 3^3 transposed convolutions. The number of encoding-decoding blocks, as well as the use of LeakyReLU activations and instance normalization layers, were chosen based on the best practices described in Isensee et al. (2018). At the same time, the network configurations that we have chosen allowed us to work with the hardware we have at hand (Titan XP 12 GB). The segmentation network outputs a 7-channel 3D volume (of the same size as the input image), corresponding to our 7 classes: background, cerebrospinal fluid (CSF), cortical gray matter (cGM), white matter (WM), deep gray matter (dGM), cerebellum and brainstem. The generator network's last convolutional layer is followed by a Tanh activation and outputs a single channel image.

For our unsupervised domain adaptation models (**Figures 3**, **4**) we used a PatchGAN discriminator as proposed in Isola et al. (2017). Its architecture consists of 5 blocks of 4^3 convolutions (with a stride of 2) with 64, 128, 256, 512, and 1 channels, respectively), instance normalization and LeakyReLU activations.

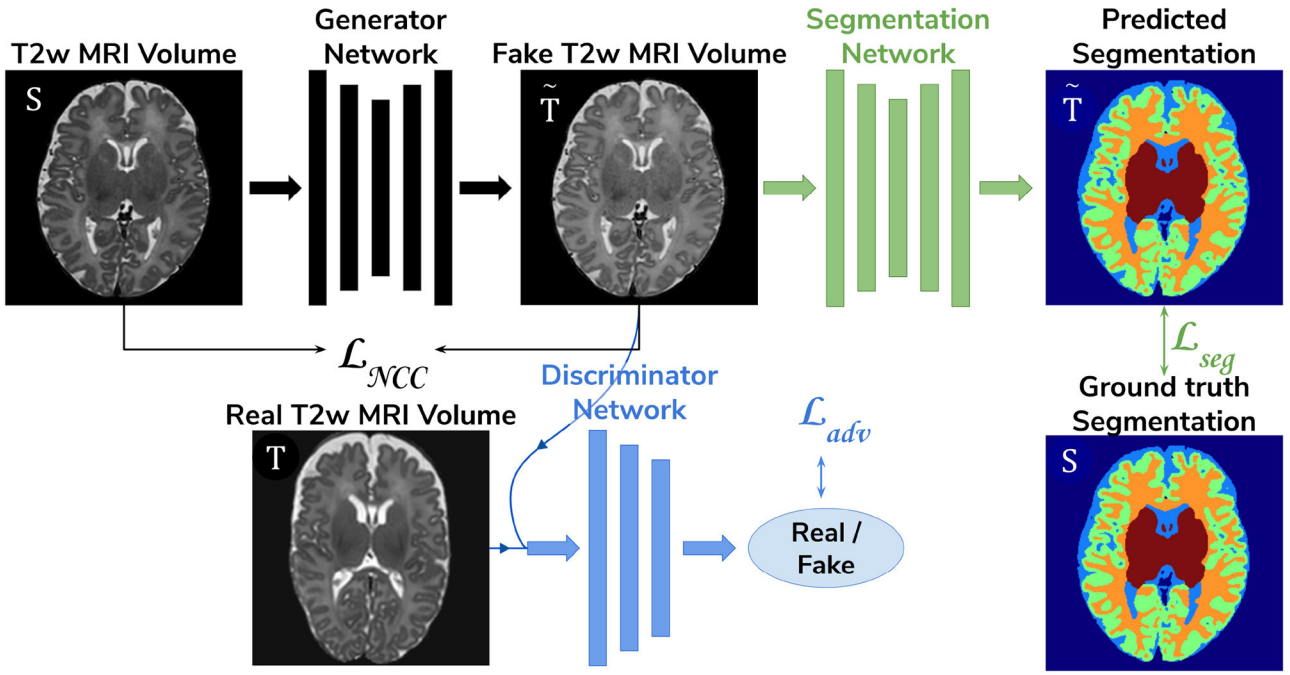


FIGURE 4 | The image space domain adaptation setup uses a generator network to produce ePrime-like T_2w MRI images (marked with $\tilde{T}$), which are then used as input into the segmentation network. The discriminator is trained to distinguish between real (ePrime) and synthesized (ePrime-like) volumes, while the generator is trained to produce realistic images in order to fool the discriminator. The NCC loss enforces image similarity between real and synthesized volumes.

The cortical parcellation network has the same architecture as the tissue segmentation network, but outputs a 12-channel 3D volume corresponding to the following cortical substructures: frontal left, frontal right, cingulate, temporal left, temporal right, insula left, insula right, parietal left, parietal right, occipital left, and occipital right, respectively. The last class represents the background.

2.4. Training

The baseline segmentation network (**Figure 2**) was trained by minimizing the generalized Dice loss (Sudre et al., 2017) between the predicted and the Draw-EM segmentation maps (Equation 1).

$$\mathcal{L}_{method_1} = \mathcal{L}_{seg} = 1 - 2\frac{\sum_{l=1}^{M} w_l \sum_n p_{ln} t_{ln}}{\sum_{l=1}^{M} w_l \sum_n p_{ln} + t_{ln}} \quad (1)$$

where $w_l = 1/(\sum_n t_{ln})^2$ is the weight of the lth tissue type, p_{ln} is the predicted probabilistic map of the lth tissue type at voxel n, t_{ln} is the target label map of the lth tissue type at voxel n, and M is the number of tissue classes. While training, we used the Adam optimizer (Kingma and Ba, 2014) with its default parameters and a decaying cyclical learning rate scheduler (Smith, 2017) with a base learning rate of $2 \cdot 10^{-6}$ and a maximum learning rate of $2 \cdot 10^{-3}$. The choice of optimizer was based on knowledge of previous image translation literature (Isola et al., 2017; Zhu et al., 2017; Liao et al., 2019; Ranzini et al., 2020) where it yielded good results. At the same time, a varying learning rate during training was shown to improve results in fewer iterations when compared to using a fixed value (Smith, 2017).

The segmentation network from the adversarial domain adaptation in the latent space model was trained to produce tissue maps on the source (dHCP) volumes. In addition, both target (ePrime) and source (dHCP) volumes were fed to the segmentation network, while the feature maps obtained from every level of its decoder arm were passed to the discriminator network which acted as a domain classifier. This was done after either up-sampling or down-sampling the feature maps to match the volume size of the second deepest layer. This model was trained by minimizing a Cross-Entropy loss between predicted and assigned target labels representing our two domains. The final loss function for our second model was therefore made up of the generalized Dice loss and an adversarial loss:

$$\mathcal{L}_{method_2} = \mathcal{L}_{seg} - \alpha \mathcal{L}_{adv} \quad (2)$$

where α was a hyperparameter increased linearly from 0 to 0.05 starting at epoch 20, and which remained equal to 0.05 from epoch 50 onward. Similar to Kamnitsas et al. (2017) we looked at the behavior of our discriminator and segmentation network when training with different values of $\alpha \in [0.02, 0.05, 0.1, 0.2, 0.5]$. We found the discriminator's accuracy during training stable for all investigated values, while the segmentation network achieved the lowest loss when $\alpha = 0.05$. The segmentation network was trained similarly to the baseline model, while the discriminator network was trained using the

TABLE 2 | Grouping of cortical substructures showing their original tissue name obtained from Draw-EM (Makropoulos et al., 2018) on the first column and their corresponding cortical subregion on the second column.

Tissue name	Cortical subregion
Anterior temporal lobe, medial part left	
Anterior temporal lobe, lateral part left	
Gyri parahippocampalis et ambiens anterior part left	
Superior temporal gyrus, middle part left	
Medial and inferior temporal gyri anterior part left	
Lateral occipitotemporal gyrus, gyrus fusiformis anterior part left	Temporal (left)
Gyri parahippocampalis et ambiens posterior part left	
Lateral occipitotemporal gyrus, gyrus fusiformis posterior part left	
Medial and inferior temporal gyri posterior part left	
Superior temporal gyrus, posterior part left	
Anterior temporal lobe, medial part right	
Anterior temporal lobe, lateral part right	
Gyri parahippocampalis et ambiens anterior part right	
Superior temporal gyrus, middle part right	
Medial and inferior temporal gyri anterior part right	
Lateral occipitotemporal gyrus, gyrus fusiformis anterior part right	Temporal (right)
Gyri parahippocampalis et ambiens posterior part right	
Lateral occipitotemporal gyrus, gyrus fusiformis posterior part right	
Medial and inferior temporal gyri posterior part right	
Superior temporal gyrus, posterior part right	
Insula left	Insula (left)
Insula right	Insula (right)
Occipital lobe left	Occipital (left)
Occipital lobe right	Occipital (right)
Cingulate gyrus, anterior part right	
Cingulate gyrus, anterior part left	Cingulate
Cingulate gyrus, posterior part right	
Cingulate gyrus, posterior part left	
Frontal lobe left	Frontal (left)
Frontal lobe right	Frontal (right)
Parietal lobe left	Parietal (left)
Parietal lobe right	Parietal (right)

Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a linearly decaying learning rate scheduler starting from $2 \cdot 10^{-3}$.

The generator network used in the image space domain adaptation approach was trained to produce synthesized ePrime volumes, while the segmentation network was trained using the same loss function, optimizer and learning rate scheduler as in the other two methods. In the previous model (adversarial domain adaptation in the latent space) we fed both dHCP and ePrime volumes to the segmentation network to obtain data agnostic feature maps. For this reason, and to allow for a fair comparison between the two unsupervised domain adaptation models, we trained the segmentation network from the image space model on both real dHCP and synthesized ePrime volumes. For both the discriminator and the generator networks the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ was used, together with a linearly decaying learning rate scheduler starting from $2 \cdot 10^{-3}$. The loss function of the discriminator was similar to that of the Least Squares GAN (Mao et al., 2017): $\mathcal{L}_D = \mathbb{E}_{x\sim T}[(D(x)-b)^2] + \mathbb{E}_{x\sim S}[(D(G(x)) - a)^2]$ where a signified the label for synthesized volumes and b was the label for real volumes. The generator and the segmentation network were trained together using the following loss:

$$\mathcal{L}_{method_3} = \mathcal{L}_{seg} + \mathcal{L}_{adv} \tag{3}$$

where $\mathcal{L}_{adv} = \mathbb{E}_{x\sim S}[(D(G(x)) - b)^2]$. An additional NCC loss was used between the real and the generated volumes in order to constrain the generator to produce realistic looking ePrime-like images. Without the additional NCC loss, the generator tends to produce images with an enlarged CSF boundary in order to match the preterm-only distribution found in the ePrime dataset, as was previously shown in Grigorescu et al. (2020).

These three methods were trained with and without data augmentation for 100 epochs, during which we used the validation sets to inform us about our models' performance and to decide on the best performing models. For data augmentation we applied: random affine transformations [with rotation angles $\theta_i \sim \mathcal{U}(-10^o, 10^o)$ and/or scaling values $s_i \sim \mathcal{U}(0.8, 1.2)$], random motion artifacts [corresponding to rotations of $\theta_i \sim \mathcal{U}(-2^o, 2^o)$ and translations of $t_i \sim \mathcal{U}(-2\,\text{mm}, 2\,\text{mm})$], and random MRI spike and bias field artifacts (Pérez-García et al., 2020). The cortical parcellation network was trained in a similar fashion as the baseline tissue segmentation network, with data augmentation in the form of random affine transformations (with the same parameters as above).

The test set was used to report our final models' results and to showcase their capability to generalize on the source domain. Finally, we produced tissue segmentation maps for all the subjects in our datasets, and used them as input into ANT's DiReCT algorithm (Tustison et al., 2013) to compute cortical thickness measures. To validate our results, we compared cortical thickness measures between subsets of the two cohorts matched for GA and PMA, for which we expect no significant difference in cortical thickness if the harmonization was successful. We also assessed the association between PMA and cortical thickness in the two cohorts.

3. RESULTS

3.1. dHCP Test Dataset

3.1.1. Baseline and Domain Adaptation Models

In our first experiment we looked at the performance of the six trained models when applied to the source (dHCP) test dataset. The aim was to assess whether our trained models were able to generalize to unseen source domain (dHCP) data for which we have reliable Draw-EM outputs. **Figure 5** summarizes the results of our trained models, showing mean Dice scores, mean Hausdorff distance calculated using SimpleITK (Lowekamp et al., 2013; Yaniv et al., 2018), precision and recall. These metrics were computed between the predicted tissue segmentation maps and the Draw-EM output labels for each of the six trained

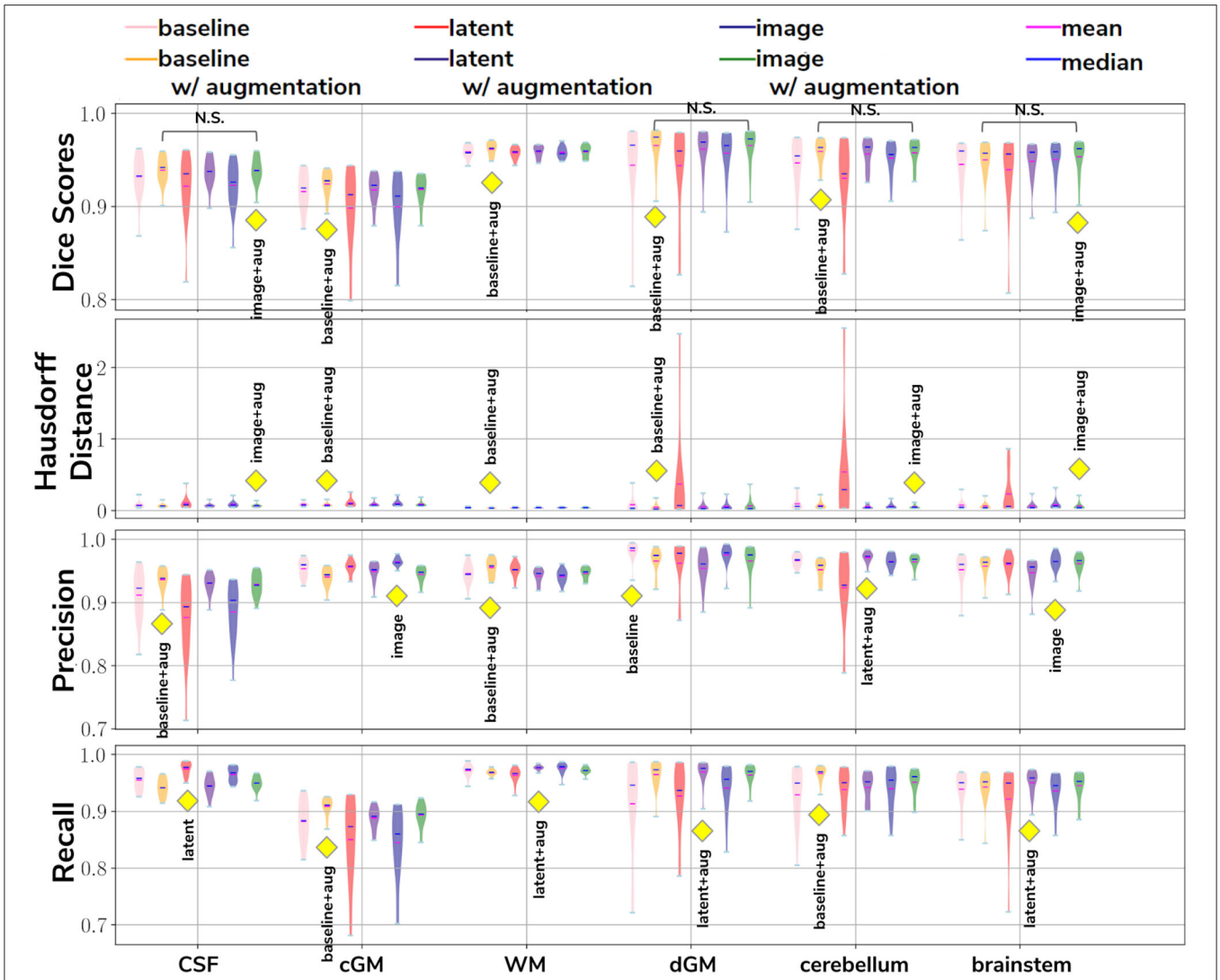


FIGURE 5 | The results on our dHCP test dataset for all six methods. The yellow diamond highlights the model which obtained the best mean score for its respective tissue type and metric. Models which obtained non-significant differences when compared to the best performing method are shown above each pair.

models. The model that obtained the best score is highlighted with the yellow diamond for each metric and tissue type. In terms of Dice scores, out of the six models, the *baseline with augmentation* and *image with augmentation* methods performed best on the source domain test dataset for CSF, dGM, cerebellum and brainstem, with no significant difference between them. For cGM and WM, the best performance was obtained by the *baseline with augmentation* model, while the domain adaptation methods showed a slight decrease in performance. The three models trained without augmentation always performed significantly worse than their augmented counterparts.

In terms of average Hausdorff distance, both the *baseline with augmentation* and *image with augmentation* models performed well, while the *latent without augmentation* model performed worse than all the other models for all tissue types. Highest precision scores were obtained by the *baseline with augmentation* model for both CSF and WM, the *image without augmentation* method for both cGM and brainstem, the *baseline without augmentation* for dGM, and the *latent with augmentation* model for cerebellum. Highest recall scores were obtained by the *baseline with augmentation* model for cGM and cerebellum, the *latent with augmentation* model for WM, dGM and brainstem, and the *latent without augmentation* model for CSF. These results show that our trained models were able to generalize to unseen source domain data, and that the performance on the dHCP dataset was not compromised by using domain adaption techniques.

3.1.2. Cortical Parcellation Network

To assess the performance of our trained cortical parcellation network, we applied it on the source (dHCP) test dataset, where the inputs were binary Draw-EM cortical gray matter tissue

TABLE 3 | Dice Scores obtained on the dHCP test set for the trained cortical parcellation network.

Tissue	Min	Max	Mean	Tissue	Min	Max	Mean
Frontal (left)	0.98	0.99	0.99	Frontal (right)	0.98	0.99	0.99
Temporal (left)	0.96	0.99	0.98	Temporal (right)	0.97	0.98	0.98
Insula (left)	0.95	0.97	0.96	Insula (right)	0.95	0.97	0.96
Parietal (left)	0.96	0.98	0.97	Parietal (right)	0.96	0.98	0.97
Occipital (left)	0.94	0.98	0.97	Occipital (right)	0.95	0.98	0.97
Cingulate	0.93	0.97	0.96				

maps. For each subject in our test dataset, the network produced a 12-channel output, consisting of: frontal left, frontal right, cingulate, temporal left, temporal right, insula left, insula right, parietal left, parietal right, occipital left, occipital right, and background, respectively. **Table 3** summarizes these results in terms of minimum, maximum and mean Dice scores for each of the 11 cortical substructures. When compared with the Draw-EM outputs (Makropoulos et al., 2018), the network obtained an overall mean Dice score of 0.97.

3.2. Validation of Data Harmonization

In order to evaluate the extent to which each of the trained models managed to harmonize the segmentation maps of the two cohorts, we looked at tissue volumes and mean cortical thickness measures between subsamples of the dHCP ($N = 30$; median GA = 30.50 weeks; median PMA = 41.29 weeks) and ePrime ($N = 30$; median GA = 30.64 weeks; median PMA = 41.29 weeks) cohort which showed comparable GA at birth and PMA at time of scan (see **Table 1**). A direct comparison between the two cohort subsets shows that the dHCP and ePrime neonates did not differ significantly in terms of sex [$\chi^2(1) < 0.001$, $p > 0.05$], or maternal ethnicity [$\chi^2(4) = 4.32$, $p > 0.05$], coded as "white or white British," "black or black British," "asian or asian British," "mixed race," and "other." As a proxy for socio-economic status, we derived an Index of Multiple Deprivation (IMD) score based on parental postcode at the time of infant birth (Department for Communities and Local Government, 2011[3]). This measure is based on seven domains of deprivation within each neighborhood compared to all others in the country: income, employment, education, skills and training, health and disability, barriers to housing and services, living environment and crime. Higher IMD values therefore indicate higher deprivation. IMD score did not differ significantly between dHCP ($M = 21.4$, $SD = 10.7$) and ePrime ($M = 18.0$, $SD = 11.6$) subsets, suggesting that these two groups are comparable in terms of environmental background.

For these two cohort subsamples with similar GA and PMA, we expected both volumes and cortical thickness measures not to differ after applying the harmonization procedures. We also investigated the relationship between PMA and volumes and cortical thickness respectively, before and after applying the harmonization. Linear regressions were performed in the comparable data subsets testing the effects of PMA and cohort on volumes (or cortical thickness), controlling for GA and sex.

3.2.1. Volumes

Figure 6 shows the tissue volumes for both the original and the predicted segmentations. Significant volume differences between the two subsamples (i.e., significant effect of cohort in the regression model) are reported above each tested model. To summarize, the *image with augmentation* model performed best, by showing no significant differences in the two cohorts for cortical gray matter, white matter, deep gray matter, cerebellum and brainstem. The cerebrospinal fluid volumes were significantly different between the two cohorts for all our trained models, as well as for the original ePrime segmentation masks.

3.2.2. Cortical Thickness

Figure 7 summarizes the results of applying the cortical thickness algorithm on the predicted segmentation maps for all six methods. Before harmonization, the matched subsets from the dHCP and ePrime cohorts showed a significant difference in mean cortical thickness [dHCP: $M = 1.73$, $SD = 0.12$; ePrime: $M = 1.93$, $SD = 0.13$; $t(58) = 6.33$, $p < 0.001$]. After applying the harmonization to the ePrime sample, mean cortical thickness no longer differed between the two subsamples for four of our methods. These results are summarized in panel H from **Figure 7**, where the models which obtained harmonized values in terms of mean cortical thickness measures are shown in bold. **Figure 7** also shows the association between PMA and mean crtical thickness before (**Figure 7A**) and after applying the models (**Figures 7B–G**) on the matched dHCP and ePrime subsets. A linear model regressing unharmonized mean cortical thickness on PMA, GA, sex, and cohort revealed a significant effect of cohort ($\beta = 0.20$; $p < 0.001$), consistent with a group difference in mean cortical thickness reported above, as well as a significant effect of PMA ($\beta = 0.04$; $p < 0.001$), consistent with an increase in cortical thickness with increasing PMA. After applying the methods, the effect of cohort was rendered non-significant for four of the methods (see highlighted panels C, E, F, G from **Figure 7**), while the effect of PMA remained stable across all six methods.

We performed a similar analysis on thickness measures of the cortical substructures. To obtain these measures, we used the original and the predicted cortical gray matter segmentation maps (obtained by applying each of our six methods) as input to the trained cortical parcellation network to predict cortical substructure masks. We then used these masks to calculate local cortical thickness measures. Our results are summarized in **Figure 8**.

3.2.3. Example Predictions

To further narrow down which of the four remaining methods was best at harmonizing our ePrime neonatal dataset, we looked at the predicted segmentations. **Figure 9** shows two example neonates from the ePrime dataset with GA = 32.9 w, PMA = 43.6 w, and with GA = 28.7 w, PMA = 44.7 w, respectively. The first column shows T_2w saggittal and axial slices, respectively, while the following four columns show example tissue prediction

[3]https://tools.npeu.ox.ac.uk/imd/

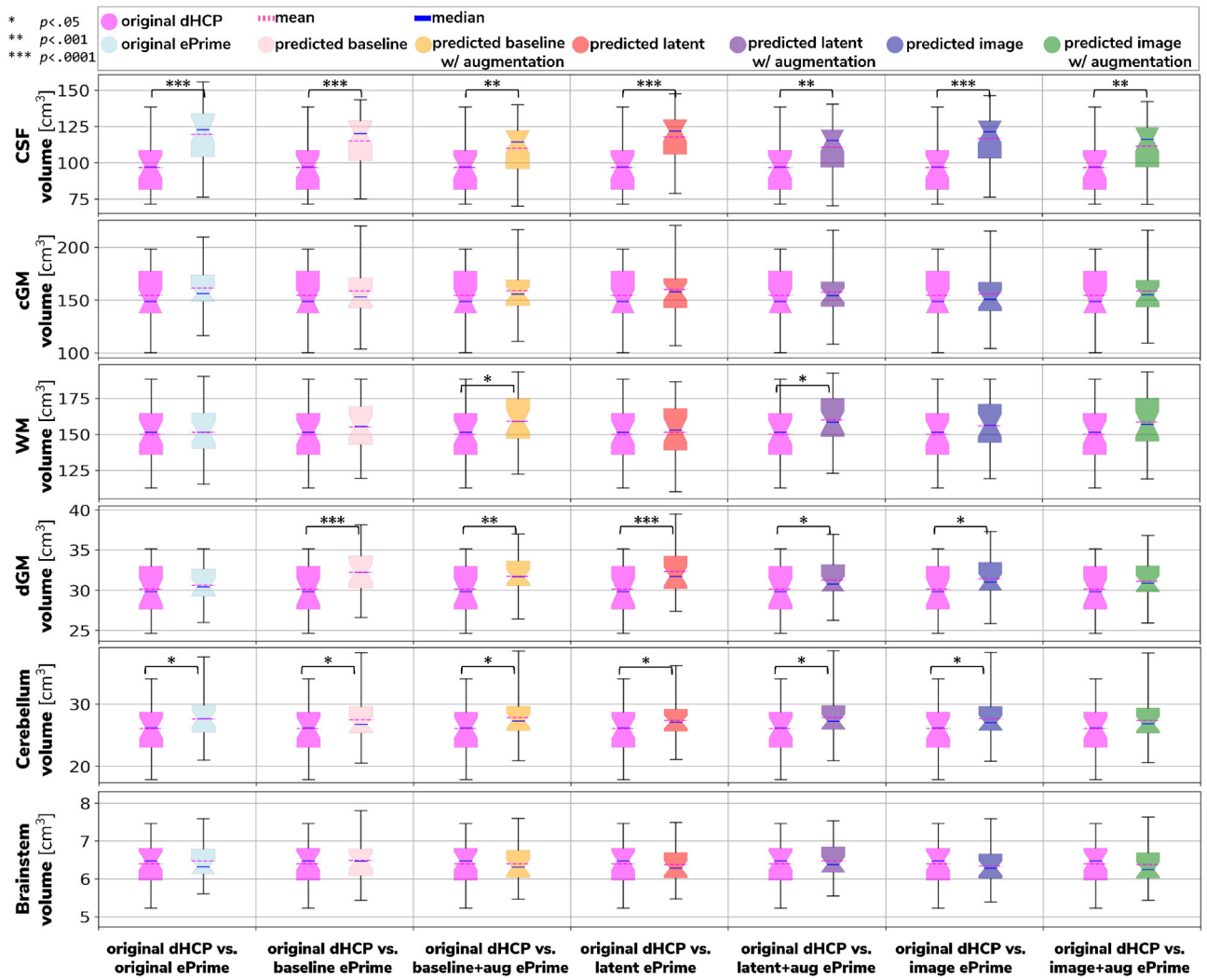


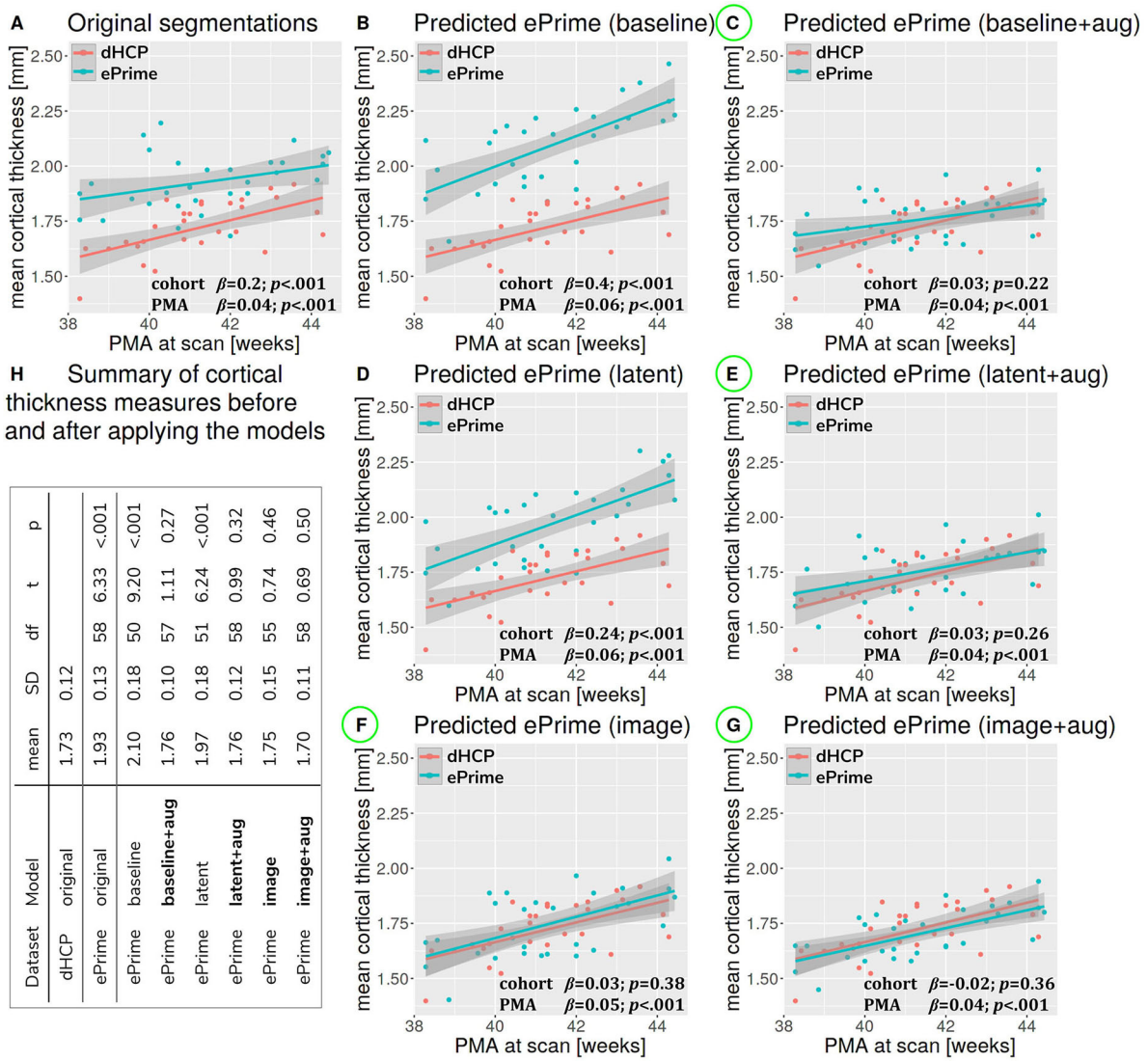
FIGURE 6 | Comparison of volume measures for our six tissue types (CSF, cGM, WM, dGM, cerebellum, and brainstem) between original Draw-EM dHCP segmentations and original Draw-EM ePrime segmentations (first column), or between original Draw-EM dHCP segmentations and ePrime segmentations obtained with the six trained models (columns 2–7). Linear regressions were performed in the comparable data subsets testing the effects of cohort on volumes, controlling for PMA, GA, and sex (volume ∼ cohort + PMA + GA + sex). The asterisks indicate a statistically significant effect of cohort in the linear regression.

maps produced by the four models: *baseline with augmentation*, *latent with augmentation*, *image*, and *image with augmentation*, respectively. Although all four methods performed well in terms of harmonizing tissue segmentation volumes and global mean cortical thickness values for the two subsamples with similar GA and PMA, previously presented quantitative results as well as the example above suggest that the *image with augmentation* method was more robust.

Finally, **Figure 10** shows the axial, sagittal and coronal slices of an ePrime neonate (GA = 32.86 w and PMA = 39.86 w). The first line shows the T_2w MR image, while the second and third lines show the CSF boundary of both the Draw-EM algorithm and the *image with augmentation* method. The green arrows point to a WM region which was misclassified by the Draw-EM pipeline as CSF. This problem was then corrected by the *image with augmentation* method.

3.3. Analysis of Harmonized Cortical Substructures

In this section we analyze the harmonized cortical gray matter segmentation maps using the *image with augmentation* model. We produce tissue segmentation maps for the entire ePrime dataset and calculate cortical thickness measures on the predicted and Draw-EM cortical gray matter tissue maps of both cohorts. In addition, we use the trained cortical parcellation network to produce cortical substructure masks. We perform a term *vs* preterm analysis on the harmonized cortical gray matter maps



Dataset	Model	mean	SD	df	t	p
dHCP	original	1.73	0.12			
ePrime	original	1.93	0.13	58	6.33	<.001
ePrime	baseline	2.10	0.18	50	9.20	<.001
ePrime	**baseline+aug**	1.76	0.10	57	1.11	0.27
ePrime	latent	1.97	0.18	51	6.24	<.001
ePrime	**latent+aug**	1.76	0.12	58	0.99	0.32
ePrime	**image**	1.75	0.15	55	0.74	0.46
ePrime	**image+aug**	1.70	0.11	58	0.69	0.50

FIGURE 7 | The association between PMA and mean cortical thickness before **(A)** and after **(B–G)** applying the data harmonization models on the matched dHCP and ePrime subsets. A linear model regressing mean cortical thickness measures on PMA, GA, sex, and cohort revealed a significant effect of cohort for the original segmentations **(A)**, and the predicted maps (**B** - *baseline without augmentation* and **D** - *latent without augmentation*). The effect of cohort was rendered non-significant for four of the methods (**C** - *baseline with augmentation*, **E** - *latent with augmentation*, **F** - *image without augmentation*, and **G** - *image with augmentation*). **(H)** summarizes cortical thickness measures before and after applying the models.

and we show the importance of harmonizing the data with a proof-of-principle application setting where we investigate the association between cortical thickness and a language outcome measure.

3.3.1. Comparison of Term and Preterm Cortical Maps

Associations between cortical thickness and GA or PMA in the full dHCP and ePrime datasets (excluding subjects with PMA > 45 weeks) for the whole cortex are depicted in **Figure 11**, where we show individual regression lines for preterm-born and term-born neonates. The first column consists of dHCP-only subjects, while the following two columns showcase both cohorts together, before and after harmonizing the cortical gray matter tissue maps.

A linear model regressing dHCP-only mean cortical thickness on PMA, GA, sex, birth weight and the interaction between PMA and GA revealed a significant effect of PMA ($\beta = 0.19$; $p < 0.001$), a significant effect of GA ($\beta = 0.16$; $p = 0.002$), and a significant effect of the interaction between PMA and GA ($\beta = -0.004$; $p = 0.002$), indicating that infants born at a lower GA showed a stronger relationship between PMA and CT. When performing the same analysis in the pooled ePrime and dHCP data before harmonizing the maps, the effect of GA and the effect of the interaction were rendered not significant (GA: $\beta = 0.009$;

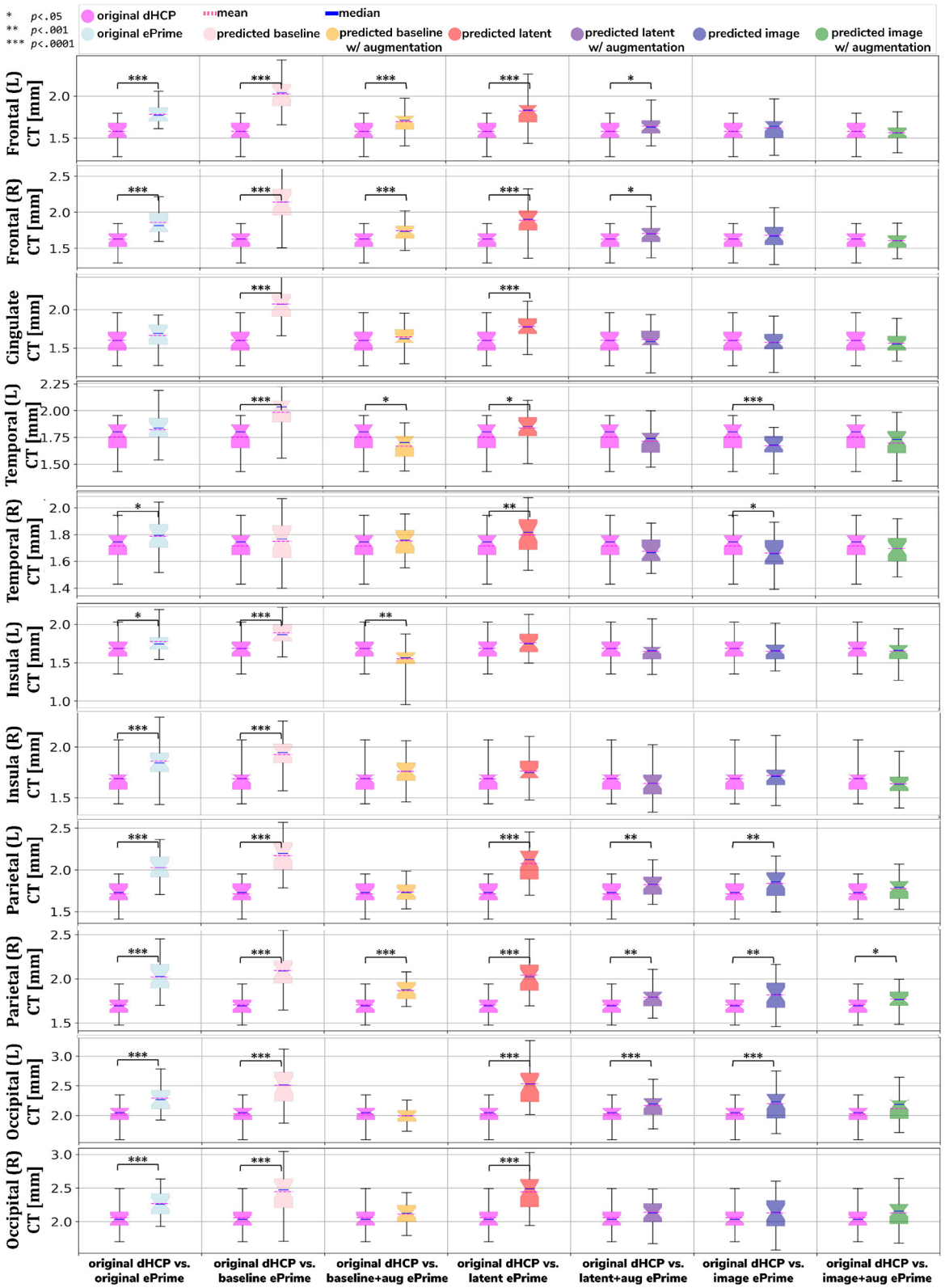


FIGURE 8 | Comparison of local mean cortical thickness measures between original Draw-EM dHCP segmentations and original Draw-EM ePrime segmentations (first column), or between original Draw-EM dHCP segmentations and ePrime segmentations obtained with the six trained models (columns 2–7). Linear regressions were performed in the comparable data subsets testing the effects of cohort on local cortical thickness measures, controlling for PMA, GA, and sex (CT ~ cohort + PMA + GA + sex). The asterisks indicate a statistically significant effect of cohort in the linear regression.

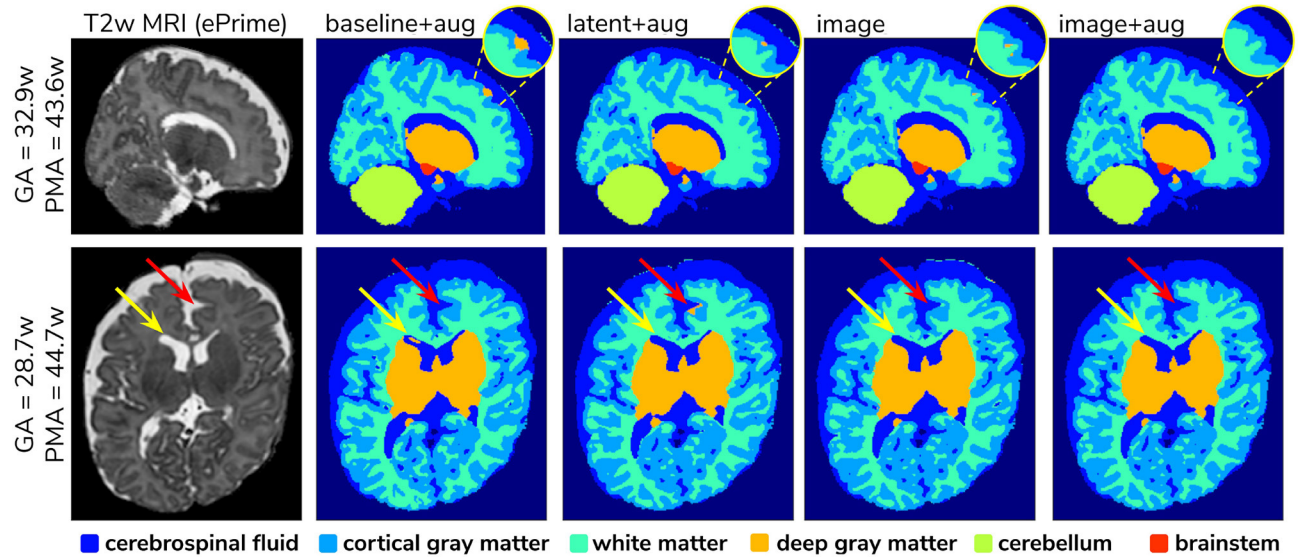


FIGURE 9 | Example predicted segmentation maps for the best performing models. On the first row we show an example where three of the models (*baseline with augmentation*, *latent with augmentation*, and *image*) misclassified a part of the cortex as being deep gray matter. This is more pronounced in the *baseline with augmentation* model, while the *latent with augmentation* and *image* show a slight improvement. The *image with augmentation* model corrected the problem entirely. On the second row the yellow arrow points to an area of CSF where the *baseline with augmentation* model misclassified it as dGM, while the other three models did not have this problem. The red arrow on the other hand points to an area where the *latent with augmentation* model misclassified cGM as deep gray matter. This problem does not appear in the other models.

$p = 0.7$ and PMA∗GA: $\beta = -0.0006$; $p = 0.5$, respectively). This is corrected after harmonizing the tissue maps, where the effects of GA ($\beta = 0.06$; $p = 0.02$) and the effects of the GA and PMA interaction ($\beta = -0.001$; $p = 0.02$) are, again, significant.

The second and third columns of **Figure 11** show that after harmonizing the tissue segmentation maps, the ePrime preterm-born neonates (green dots) are brought downwards into a comparable range of values to the dHCP preterms (red dots). Moreover, when plotting the cortical thickness measures against PMA, after harmonizing the tissue maps, the intersection between the two individual regression lines (term and preterm-born neonates) happens at roughly the same age (PMA = 38.5 weeks) as in the dHCP-only dataset.

We extended the term *vs* preterm analysis on cortical thickness substructures. **Figure 12** shows the results of applying a linear model regressing mean cortical thickness measures on PMA, GA, sex, birth weight and prematurity, where significant differences ($p < 0.05$) between the two cohorts (term and preterm-born neonates) are highlighted in the image.

3.3.2. Behavioral Outcome Association

As a final proof-of-principle, we demonstrate the importance of data harmonization in an application setting investigating the association between neonatal cortical thickness and a behavioral outcome measure. For this, we consider language abilities as assessed between 18 and 24 months in both dHCP and ePrime cohorts using the Bayley Scales of Infant and Toddler Development (Bayley, 2006). Age-normed composite language scores were available for 203 toddlers from the dHCP cohort (M = 96.43; SD = 14.89) and 136 toddlers from the ePrime cohort (M = 91.25; SD = 17.37). For the neonatal cortical thickness measure, we focus on the left and right frontal cortex for illustration.

Regressing composite language score against left or right frontal cortical thickness in each cohort separately, controlling for PMA, GA, sex and intracranial volume showed that there was no significant association between neonatal left/right frontal cortical thickness and language abilities at toddler age in either of the cohorts. However, when pooling data from both cohorts together and rerunning the same analysis (using un-harmonized cortical thickness measures), a significant association between left/right frontal cortical thickness and language abilities is seen (left: $\beta = -17.56$, $p < 0.05$, right: $\beta = -18.76$, $p < 0.05$), suggesting that greater frontal cortical thickness at term-equivalent age is associated with reduced language abilities at toddler age.

However, as can be seen in **Figure 13**, this is likely a spurious effect due to (artifactually) heightened cortical thickness values in un-harmonized ePrime data combined with lower language composite scores in the ePrime cohort (consistent with effects typically observed in preterm cohorts). Indeed, when rerunning the same analysis on harmonized data pooled across both cohorts, the effect of cortical thickness on language ability is rendered non-significant in both left ($\beta = -13.99$, $p = 0.15$) and right ($\beta = -16.69$, $p = 0.068$) frontal cortex, consistent with the ground-truth findings in each individual cohort.

4. DISCUSSION AND FUTURE WORK

In this paper we studied the application and viability of unsupervised domain adaptation methods for harmonizing

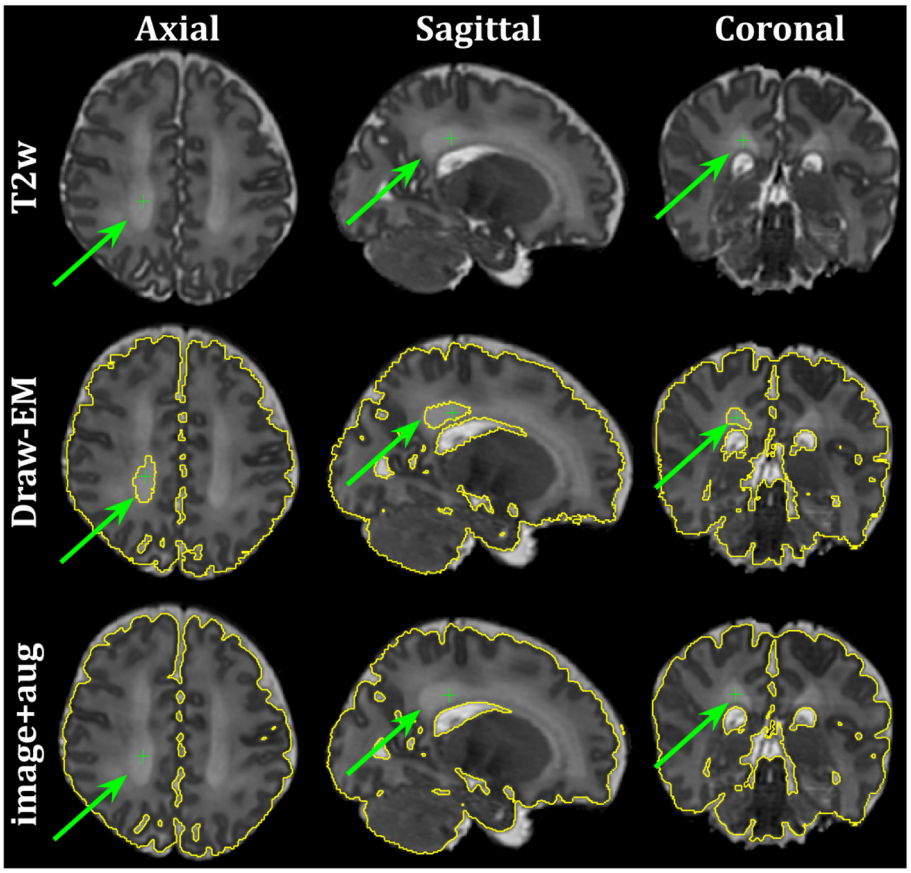


FIGURE 10 | Example of a neonate from the ePrime dataset with GA = 32.86 w and PMA = 39.86 w where the Draw-EM algorithm performed worse than our proposed *image with augmentation* model. The green arrow points at a region which was segmented as CSF by Draw-EM, but then corrected by our model.

tissue segmentation maps of two neonatal datasets (dHCP and ePrime). Our aim was to obtain volumetric and cortical thickness measures that are only affected by brain anatomy and not by the acquisition protocol or scanner, in order to improve the statistical power of imaging or imaging-genetic studies. We proposed an image-based domain adaptation model where a tissue segmentation network was trained with real dHCP and synthesized ePrime T_2w 3D MRI volumes. The generator network was trained to produce realistic images in order to fool a domain discriminator, while also minimizing an NCC loss which aimed to enforce image similarity between real and synthesized images (Grigorescu et al., 2020). We trained this model using dHCP Draw-EM segmentation maps, and we compared it with a baseline 3D U-Net (Çiçek et al., 2016), and a latent space domain adaptation method (Kamnitsas et al., 2017). The three methods were trained with and without data augmentation (Pérez-García et al., 2020).

First, we looked at the performance of each of the six trained models on the source (dHCP) test dataset, by comparing predicted tissue segmentation maps with the Draw-EM output labels, with the aim of measuring fidelity of our trained segmentation methods for the original dHCP domain. Our results on the source (dHCP) test dataset suggest that our trained models were able to generalize to unseen source domain data. At the same time, Dice score results on the test set for the proposed *image with augmentation* model are high and are similar in performance when compared with the *baseline with augmentation* method. This suggests that adding the contrast transfer step does not diminish the quality of the segmentations.

We then analyzed the extent to which each of the 6 trained models managed to harmonize the tissue segmentation maps of our two cohorts, by looking at tissue volumes and mean cortical thickness measures between subsamples of the dHCP and ePrime cohorts which showed comparable GA at birth and PMA at time of scan, as well as similar gender and maternal ethnicity. Our results showed that our proposed model (*image with augmentation*) harmonized the predicted tissue segmentation maps in terms of cortical gray matter, white matter, deep gray matter, cerebellum and brainstem volumes (**Figure 6**). In terms of mean global cortical thickness measures, four of the trained methods (*baseline with augmentation*, *latent with augmentation*, *image*, and *image with augmentation*) achieved comparable values when compared to the dHCP subset. In fact, we hypothesize that these four methods provided the best overall results because either they were trained using data augmentation or they acted as a deep learning-based augmentation technique

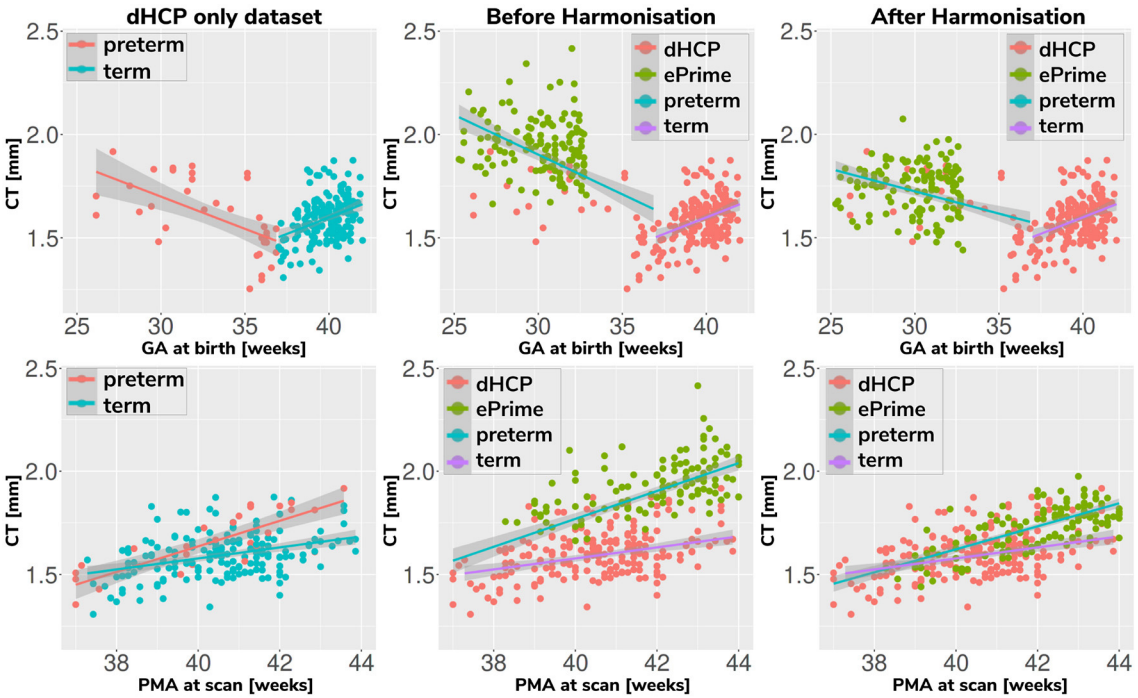
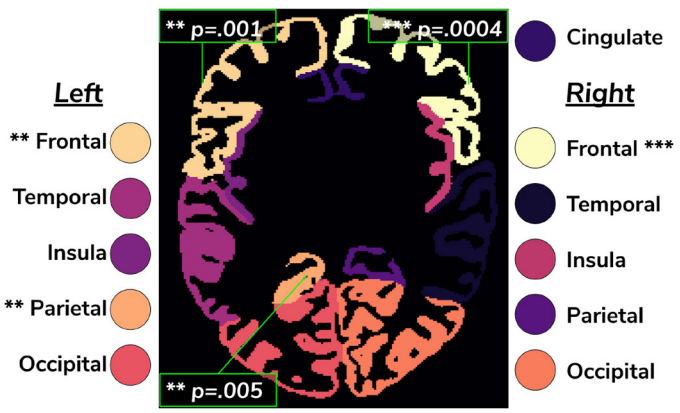


FIGURE 11 | Mean cortical thickness measures in our dHCP dataset (first column), and in both cohorts before (second column) and after (third column) harmonizing the tissue segmentation maps. The first row plots the cortical thickness measures against GA, while the second row plots the cortical thickness measures against PMA, with individual regression lines on top.



Comparison of cortical thickness measures between term and preterm-born neonates

Tissue	*p-val*	Tissue	*p-val*
Global	**0.009**	Cingulate	0.74
Frontal (left)	**0.001**	**Frontal (right)**	**0.0004**
Temporal (left)	0.270	Temporal (right)	0.827
Insula (left)	0.376	Insula (right)	0.093
Parietal (left)	**0.005**	Parietal (right)	0.488
Occipital (left)	0.058	Occipital (right)	0.052

FIGURE 12 | Comparison of cortical thickness measures for the whole cortex and for each of the 11 cortical subregions between term and preterm-born neonates. The results of the linear regression are reported in the table in terms of differences between term and preterm-born neonates.

(Sandfort et al., 2019), which made the segmentation network more robust to the different contrast, population bias and acquisition protocol of the ePrime dataset.

Using the cortical parcellation network, we also produced cortical thickness measures for the 11 cortical subregions (see **Table 2**). Again, the models trained with augmentation performed better than their no augmentation counterparts (see **Figure 8**). However, our proposed *image with augmentation* model performed best, whereby ePrime values, tending toward higher values before harmonization, were brought downwards into a comparable range of values to dHCP, for 10 out of 11 cortical subregions (see **Figure 8** last column). For the right parietal lobe, our proposed method outperformed the original segmentations and the other 5 models, but did not manage to bring the values down to a non-significant range. One potential reason for this is that, on a visual inspection, the ePrime cohort appears to suffer from more partial volume artifacts than its dHCP counterpart, which can confuse the segmentation

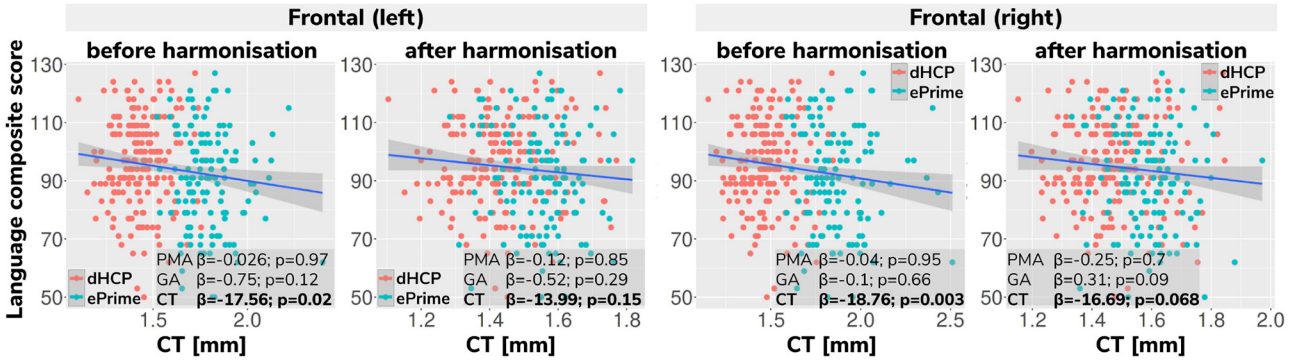


FIGURE 13 | Language composite score against predicted left and right frontal cortical thickness measures before and after harmonizing the tissue segmentation maps. Without harmonization (columns 1 and 3) there appears to be a significant association between left or right frontal cortical thickness and language abilities, but after harmonization (columns 2 and 4) the effect of cortical thickness on language ability is rendered non-significant in both left and right frontal cortex. This demonstrates the importance of data harmonization without which pooling images from separate datasets can lead to spurious findings that are driven by differences in acquisitions rather than by true underlying effects.

network and can lead to overestimation of the cortical gray matter/cerebrospinal fluid boundary.

A close inspection of the predicted tissue segmentation maps (see **Figure 9**) also showed that our proposed model (*image with augmentation*) corrected misclassified voxels which were prevalent in the other 3 methods. At the same time, the proposed *image with augmentation* method outperformed the original Draw-EM segmentation by correcting a region of WM which was wrongly classified as CSF (see **Figure 10**). Our results suggest that, in terms of consistency of volumes and regional cortical thickness measures derived from dHCP and ePrime neonates (**Figures 6**, **8**), as well as the qualitative examples (**Figures 9**, **10**), our proposed *image with augmentation* model resulted in more consistent outputs than the other methods.

We used the harmonized cortical segmentation maps to look at differences in both global and local cortical thickness measures between term and preterm-born neonates. We showed in **Figure 12** that our harmonized cortical gray matter maps resulted in global thickness measures which were comparable with the dHCP-only neonates, while also revealing a significant effect of GA and the interaction between age at scan and at birth. We performed a similar analysis on the local cortical thickness measures and highlighted three regions of interest (frontal left, frontal right, and parietal left) which showed significant differences between the two cohorts (see **Figure 12**). These regions are consistent with previous studies (Nagy et al., 2011) where cortical thickness measures were shown to differ in preterm-born neonates when compared to term-born neonates in an adolescent cohort.

Finally, we showed the importance of harmonizing the cortical tissue maps by investigating the association between neonatal cortical thickness and a language outcome measure. After harmonization, regressing language composite score against predicted left or right frontal cortical thickness in the two pooled datasets, showed no significant effect of cortical thickness (second column of **Figure 13**), consistent with the ground-truth results seen in each cohort individually. This analysis demonstrates that without data harmonization, pooling images from separate datasets can lead to spurious findings that are driven by systematic differences in acquisitions rather than by true underlying effects. Our harmonization allows for our two datasets to be combined into joint analyses while preserving the underlying structure of associations with real-world outcomes.

Our study was focused on single-source unsupervised domain adaptation approaches, which might limit application in terms of applying the method to a different neonatal dataset. However, by utilizing reliable tissue segmentation maps from multiple neonatal databases, the proposed model can be extended to a multi-source domain adaptation pipeline (Mansour et al., 2008; Xu et al., 2018). Additionally, the latent based domain adaptation method was trained using the features at every layer of the decoding branch, without analyzing different combinations of the encoding-decoding layers. Future work will therefore aim to systematically evaluate our design choices via ablation studies. At the same time, we focused our work on investigating structural (T_2w) datasets only, and in future we aim to extend this study to harmonize diffusion data as well.

DATA AVAILABILITY STATEMENT

The code developed for this study are available online on Github (https://github.com/irinagrigorescu). Imaging data collected for the dHCP are available in early 2021 at http://developingconnectome.org/. Requests for Data Sharing for the ePrime dataset should be made to the Chief Investigator of (Edwards et al., 2018) doi: 10.1136/archdischild-2017-313102.

ETHICS STATEMENT

The studies involving human participants were approved by the National Research Ethics Committee (dHCP, REC: 14/Lo/1169; EPrime, REC: 09/H0707/98). Informed written consent was given by parents prior to scanning.

AUTHOR CONTRIBUTIONS

IG prepared the manuscript, implemented the code for the domain adaptation models and the analysis. LV participated in the implementation of the analysis code, the study design and interpretation of the results. AU assisted with data preprocessing, design of the study and interpretation of the results. DB performed preprocessing of the dHCP and ePrime datasets. LC-G developed MRI acquisition protocols for the neonatal dHCP datasets. CN participated in the study design and interpretation of the results. ADE and JVH are coordinators of the dHCP project. MM supervised all stages of the current research. MD conceptualized the study, supervised all stages of the current research and preparation of the manuscript. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

FUNDING

This work was supported by the Academy of Medical Sciences Springboard Award [SBF004\1040], Medical Research Council (Grant nos. [MR/K006355/1] and [MR/S026460/1]), European Research Council under the European Union's Seventh Framework Programme [FP7/20072013]/ERC grant agreement no. 319456 dHCP project, the EPSRC Research Council as part of the EPSRC DTP (grant Ref: [EP/R513064/1]), the Wellcome/EPSRC Centre for Medical Engineering at King's College London [WT 203148/Z/16/Z], the NIHR Clinical Research Facility (CRF) at Guy's and St Thomas', and by the National Institute for Health Research Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The EPrime study was funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (grant reference no. [RP-PG-0707-10154]).

ACKNOWLEDGMENTS

We thank everyone who was involved in acquisition and analysis of the datasets. We thank all participants and their families. This paper is an extension of our previous work (Grigorescu et al., 2020).

REFERENCES

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development*. PsychCorp; Pearson. San Antonio, TX: The Psychological Corporation.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens; Cham: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Cordero-Grande, L., Hughes, E. J., Hutter, J., Price, A. N., and Hajnal, J. V. (2018). Three-dimensional motion corrected sensitivity encoding reconstruction for multi-shot multi-slice MRI: application to neonatal brain imaging. *Magn. Reson. Med.* 79, 1365–1376. doi: 10.1002/mrm.26796

Edwards, A. D., Redshaw, M. E., Kennea, N., Rivero-Arias, O., Gonzales-Cinca, N., Nongena, P., et al. (2018). Effect of mri on preterm infants and their families: a randomised trial with nested diagnostic and economic evaluation. *Arch. Dis. Childhood Fetal Neonatal Ed.* 103, F15–F21. doi: 10.1136/archdischild-2017-313102

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by backpropagation," in *International Conference on machine learning. Proceedings of the 32nd International Conference on Machine Learning,* Vol. 37, 1180–1189.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., et al. (2017). "Transfer learning for domain adaptation in MRI: application in brain lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC; Cham: Springer), 516–524. doi: 10.1007/978-3-319-66179-7_59

Grigorescu, I., Cordero-Grande, L., Batalle, D., Edwards, A. D., Hajnal, J. V., Modat, M., et al. (2020). "Harmonised segmentation of neonatal brain MRI: a domain adaptation approach," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis* (Cham: Springer International Publishing), 253–263.

Hughes, E. J., Winchman, T., Padormo, F., Teixeira, R., Wurie, J., Sharma, M., et al. (2017). A dedicated neonatal brain imaging system. *Magn. Reson. Med.* 78, 794–804. doi: 10.1002/mrm.26462

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., et al. (2018). nnu-net: self-adapting framework for u-net-based medical image segmentation. *arXiv preprint* arXiv:1809.10486. doi: 10.1007/978-3-658-25326-4_7

Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., et al. (2017). "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Information Processing in Medical Imaging* (Cham: Springer International Publishing).

Kerfoot, E., Puyol-Antón, E., Ruijsink, B., Ariga, R., Zacur, E., Lamata, P., et al. (2019). "Synthesising images and labels between MR sequence types with CycleGAN," in *Domain Adaptation and Representation Transfer and Medical Image Learning With Less Labels and Imperfect Data* (Shenzhen; Cham: Springer). doi: 10.1007/978-3-030-33391-1_6

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]* arXiv:1412.6980.

Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M. A., Hajnal, J. V., and Schnabel, J. A. (2012). Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Med. Image Anal.* 16, 1550–1564. doi: 10.1016/j.media.2012.07.004

Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., et al. (2019). Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Sci. Rep.* 9, 1–15. doi: 10.1038/s41598-019-43299-z

Liao, H., Lin, W.-A., Zhou, S. K., and Luo, J. (2019). Adn: artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Trans. Med. Imaging* 39, 634–643. doi: 10.1109/TMI.2019.2933425

Lowekamp, B., Chen, D., Ibáñez, L., and Blezek, D. (2013). The design of SimpleITK. *Front. Neuroinform.* 7:45. doi: 10.3389/fninf.2013.00045

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., et al. (2018). The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage* 173, 88–112. doi: 10.1016/j.neuroimage.2018.01.054

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2008). Domain adaptation with multiple sources. *Adv. Neural Inform. Process. Syst.* 21, 1041–1048.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., and Paul Smolley, S. (2017). "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* 19, 1236–1246. doi: 10.1093/bib/bbx044

Nagy, Z., Lagercrantz, H., and Hutton, C. (2011). Effects of preterm birth on cortical thickness measured in adolescence. *Cereb. Cortex* 21, 300–306. doi: 10.1093/cercor/bhq095

Orbes-Arteaga, M., Varsavsky, T., Sudre, C. H., Eaton-Rosen, Z., Haddow, L. J., Sørensen, L., et al. (2019). "Multi-domain adaptation in brain MRI through paired consistency and adversarial learning," in *Domain Adaptation and Representation Transfer and Medical Image Learning With Less Labels and Imperfect Data*, eds Q. Wang, F. Milletari, H. V. Nguyen, S. Albarqouni, M. J. Cardoso, N. Rieke, Z. Xu, K. Kamnitsas, V. Patel, B. Roysam, S. Jiang, K. Zhou, K. Luu, and N. Le (Cham: Springer International Publishing), 54–62.

Pérez-García, F., Sparks, R., and Ourselin, S. (2020). TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv [Preprint]* arXiv:2003.04696.

Ranzini, M. B. M., Groothuis, I., Kläser, K., Cardoso, M. J., Henckel, J., Ourselin, S., et al. (2020). "Combining multimodal information for metal artefact reduction: an unsupervised deep learning framework," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (Iowa City, IA), 600–604. doi: 10.1109/ISBI45749.2020.9098633

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imaging* 18, 712–721. doi: 10.1109/42.796284

Sandfort, V., Yan, K., Pickhardt, P. J., and Summers, R. M. (2019). Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Sci. Rep.* 9, 1–9. doi: 10.1038/s41598-019-52737-x

Schuh, A., Makropoulos, A., Robinson, E. C., Cordero-Grande, L., Hughes, E., Hutter, J., et al. (2018). Unbiased construction of a temporally consistent morphological atlas of neonatal brain development. *bioRxiv.* 251512. doi: 10.1101/251512

Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., et al. (2017). Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. *Am. J. Neuroradiol.* 38, 1501–1509. doi: 10.3174/ajnr.A5254

Smith, L. N. (2017). "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE), 464–472.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Lecture Notes Comput. Sci.* doi: 10.1007/978-3-319-67558-9_28

Takao, H., Hayashi, N., and Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. *J. Magn. Reson. Imaging* 34, 438–444. doi: 10.1002/jmri.22636

Tustison, N. J., Avants, B. B., Cook, P. A., Song, G., Das, S., van Strien, N., et al. (2013). "The ANTs cortical thickness processing pipeline," in *Medical Imaging 2013: Biomedical Applications in Molecular, Structural, and Functional Imaging* (Lake Buena Vista, FL: PIE Medical Imaging). doi: 10.1117/12.2007128

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. *arXiv [Preprint]* arXiv:1607.08022.

Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. (2018). "Deep cocktail network: multi-source unsupervised domain adaptation with category shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 3964–3973.

Yaniv, Z., Lowekamp, B. C., Johnson, H. J., and Beare, R. (2018). SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. *J. Digit. Imaging* 31, 290–303. doi: 10.1007/s10278-017-0037-8

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 2223–2232.

Disclaimer: The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 15 June 2021
doi: 10.3389/fnins.2021.646034

Bundle-Specific Axon Diameter Index as a New Contrast to Differentiate White Matter Tracts

Muhamed Barakovic[1,2,3,4†], Gabriel Girard[1,5,6†], Simona Schiavi[1,7], David Romascano[1], Maxime Descoteaux[8], Cristina Granziera[3,4], Derek K. Jones[2,9,10], Giorgio M. Innocenti[1,11,12‡], Jean-Philippe Thiran[1,5,6†] and Alessandro Daducci[7†]*

[1] Signal Processing Lab 5, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [2] Cardiff University Brain Research Imaging Centre, Cardiff University, Cardiff, United Kingdom, [3] Translational Imaging in Neurology (ThINk) Basel, Department of Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland, [4] Neurologic Clinic and Policlinic, Departments of Medicine, Clinical Research and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland, [5] CIBM Center for BioMedical Imaging, Lausanne, Switzerland, [6] Radiology Department, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland, [7] Department of Computer Science, University of Verona, Verona, Italy, [8] Sherbrooke Connectivity Imaging Lab, Université de Sherbrooke, Sherbrooke, QC, Canada, [9] Neuroscience and Mental Health Research Institute, Cardiff University, Cardiff, United Kingdom, [10] Mary MacKillop Institute for Health Research, Australian Catholic University, Melbourne, VIC, Australia, [11] Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden, [12] Brain and Mind Institute, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

In the central nervous system of primates, several pathways are characterized by different spectra of axon diameters. *In vivo* methods, based on diffusion-weighted magnetic resonance imaging, can provide axon diameter index estimates non-invasively. However, such methods report voxel-wise estimates, which vary from voxel-to-voxel for the same white matter bundle due to partial volume contributions from other pathways having different microstructure properties. Here, we propose a novel microstructure-informed tractography approach, $\text{COMMIT}_{\text{AxSize}}$, to resolve axon diameter index estimates at the streamline level, thus making the estimates invariant along trajectories. Compared to previously proposed voxel-wise methods, our formulation allows the estimation of a distinct axon diameter index value for each streamline, directly, furnishing a complementary measure to the existing calculation of the mean value along the bundle. We demonstrate the favourable performance of our approach comparing our estimates with existing histologically-derived measurements performed in the corpus callosum and the posterior limb of the internal capsule. Overall, our method provides a more robust estimation of the axon diameter index of pathways by jointly estimating the microstructure properties of the tissue and the macroscopic organisation of the white matter connectivity.

Keywords: human brain, white-matter axon signature, diffusion MRI, tractography, microstructure, microstructure informed tractography

OPEN ACCESS

Edited by:
Tim B. Dyrby, Technical University of Denmark, Denmark

Reviewed by:
Viktor Vegh, The University of Queensland, Australia
Ivana Drobnjak, University College London, United Kingdom

****Correspondence:***
Muhamed Barakovic
muhamed.barakovic@epfl.ch

†These authors have contributed equally to this work

‡Dedicated
This paper to our friend and colleague Giorgio M. Innocenti, who passed away on January 12, 2021

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *24 December 2020*
Accepted: *17 May 2021*
Published: *15 June 2021*

Citation:
Barakovic M, Girard G, Schiavi S, Romascano D, Descoteaux M, Granziera C, Jones DK, Innocenti GM, Thiran J-P and Daducci A (2021) Bundle-Specific Axon Diameter Index as a New Contrast to Differentiate White Matter Tracts. Front. Neurosci. 15:646034. doi: 10.3389/fnins.2021.646034

1. INTRODUCTION

The *white matter of the central nervous system* comprises axons with different diameters (Peters et al., 1991) organized in pathways, tracts, bundles or fascicles. Diameters correlate with: (i) the size of the parent cell body (Tomasi et al., 2012); (ii) the size and density of synaptic boutons (Innocenti and Caminiti, 2017); (iii) conduction velocity (Hursh, 1937), which together with axon length determines conduction delays between brain sites; and possibly, (iv) the frequency of firing (Perge et al., 2012). Being able to quantify and characterise these different aspects may be critical to understanding sensory, motor, and cognitive functions. In particular, as the axon diameter is strictly related to conduction velocity (Ritchie, 1982; Drakesmith et al., 2019), it is associated with the flow of information between different cortical sites and is thus a critical feature when trying to understand the relationship between the structural and functional connectivity of the brain (Honey et al., 2010). Reliable estimates of axon diameter are also of utmost importance for interpreting pathological cases (DeLuca et al., 2004; Zikopoulos and Barbas, 2013; Huang et al., 2016).

First attempts to characterize the composition of neuronal pathways in the central nervous system used *histological techniques* (Aboitiz et al., 1992; Tomasi et al., 2012; Innocenti et al., 2018) and focused on samples of animal tissue. Besides being possible only *ex vivo*, these analyses require laborious measurements of axon diameters in a few slices along the course of known pathways. Per contra, *diffusion-weighted magnetic resonance imaging* (DW-MRI) is a non-invasive technology that can provide *in vivo* structural information on white matter pathways by probing the motion of water molecules and analyzing how it is influenced by the cellular structure of the tissue (Le Bihan and Breton, 1985; Moseley et al., 1990; Beaulieu and Allen, 1994). Compared to histological measurements, this technology is faster and non-invasive. Therefore, it can be applied to the living human brain, with enormous potential in terms of information that can be recovered.

On the one hand, it is possible to estimate the course of major pathways using *tractography*; for a review, see (Jeurissen et al., 2017) and references therein. These fiber-tracking methods approximate the macroscopic trajectory of axons by seeking pathways of maximum coherence of estimates of fibre orientation derived in each voxel from DW-MRI. Each reconstructed trajectory, or streamline, represents a coherent set of axons coursing together. Despite a large number of algorithms developed, none of the existing methods can provide information about the axon diameter of the individual reconstructed fiber bundles, as tractography only reconstructs their macroscopic trajectory. On the other hand, a variety of DW-MRI biophysical models have been proposed in the literature to obtain such information at the voxel level. Pioneering work in this field was done by Assaf et al. (2008), who proposed a method to estimate *axon diameter distributions* on an *ex vivo* spinal cord sample, exploiting the simple organization of the tissue with axons having a single, known orientation. Their model, AxCaliber, was later employed to study *in vivo* the axon composition of the corpus callosum in rodents (Barazany et al., 2009). A major limitation is that the DW-MRI signal must be acquired perpendicular to the axons main orientation and, hence, it requires prior knowledge on the orientation of the bundle to study. The ActiveAx technique developed by Alexander et al. (2010) removed this constraint by probing the DW-MRI signal along multiple directions and estimating orientationally-invariant features of the axons, thus not requiring any prior knowledge on their orientation. ActiveAx extended axon diameter index estimation to the whole brain but at the price of providing estimates of the *mean axon diameter* rather than the full distribution. The model was validated in monkeys and humans (Alexander et al., 2010; Dyrby et al., 2013), *in vivo* and *ex vivo*, and the estimated *trend* of the mean diameters in the corpus callosum agreed with histology.

Despite their attractiveness, current techniques for axon diameter estimation with DW-MRI suffer from several *fundamental limitations* which render them unsuitable for estimating conduction velocity and connectomics studies in the whole brain. First, the estimation is performed voxel-wise and independently in each imaging voxel, neglecting the fact that axons are continuous three-dimensional structures that are not limited to the extent of the voxel. This makes it impossible to infer the full course of the axons passing through that location or whether the estimated values correspond to distinct fiber bundles. Second, most methods implicitly assume a single axon population inside a voxel and cannot cope with complex fiber configurations such as crossing and fanning. In such voxels, [estimated to be as high as 90 % of all white-matter voxels (Jeurissen et al., 2013)], the models provide biased estimates as they suffer from severe overestimation of the axon diameters (Alexander et al., 2010), limiting de facto their applicability to specific areas of the brain, e.g., mid-sagittal plane of the corpus callosum. Recent advances extended these models to multiple fiber populations (Barazany et al., 2011; Zhang et al., 2011a; Auria et al., 2015; Farooq et al., 2016) and orientation dispersion (Zhang et al., 2011b), allowing for a more accurate estimation in complex fiber configurations. Although these methods showed consistent differences in the axon diameter index estimation from various axonal bundles, they remain limited to voxel-wise estimates, and are unable to recover bundle-specific methods. It would be desirable to obtain an accurate estimation along bundle trajectories, and in all white matter voxels, allowing for the characterization of the axon composition of individual fiber bundles. Lastly, the accuracy of the estimates crucially depends on the strength of the diffusion gradients that can be generated by the MRI scanners (Dyrby et al., 2013; Nilsson et al., 2017; Jones et al., 2018; Huang et al., 2020; Paquette et al., 2020) and other acquisition protocol (Gore et al., 2010; Siow et al., 2013; Kakkar et al., 2018; Xu et al., 2014, 2016; Drobnjak et al., 2016; Fan et al., 2020; Veraart et al., 2020), which affects the accuracy of the parameters as well (Drobnjak et al., 2010; Harkins et al., 2021); conventional human scanners are equipped with gradient systems up to 80 mT m^{-1}, which do not provide the required sensitivity to axon diameters (Novikov et al., 2018; Veraart et al., 2020).

In this paper, we propose a novel method to overcome the above limitations and enable, for the first time, *a non-invasive characterization of an invariant value of axon diameter index per streamline* in the living human brain. Our method

combines tractography with a microstructure model of the neuronal tissue and uses DW-MRI data acquired with a 3 T Connectom scanner capable of exploiting diffusion gradients up to 300 mT m^{-1}. We demonstrate the favourable performance of our method comparing our estimates with existing histologically-derived measurements (Caminiti et al., 2009) performed in the corpus callosum and the posterior limb of the internal capsule. Estimating bundle-specific axon diameter index within each voxel of the whole white matter would represent a major advance in neuroscience, as this could shed more light on the relation between structural and functional connectivity (Honey et al., 2010) and improve our understanding of brain dynamics.

2. THEORY AND BACKGROUND

2.1. Voxel-Wise vs. Bundle-Specific Axon Diameter Estimation

To illustrate the importance of bundle-specific axon diameter index estimation, let us consider the simple example in **Figure 1**. This synthetic dataset consists of two crossing fiber populations characterized by different axon compositions, with the green bundle containing larger axons than the blue one (**Figure 1A**). Today, the axon diameter index of a bundle is characterized using *tractometry*. This procedure indirectly approximates bundle-specific statistics by first estimating the axon diameter index with voxel-wise techniques in every voxel of the image (**Figure 1B**). For simplicity, we report only the estimated mean values rather than the full distributions. Then, the representative value of such a metric, for a given bundle, is obtained by *averaging these values* in all the voxels that are traversed by the streamlines belonging to the bundle. The purpose of this work is to develop a novel technique capable of estimating *bundle-specific statistics*, thus allowing us to obtain more reliable estimates of its axon composition (**Figure 1D**).

2.2. Microstructure-Informed Tractography

Even though DW-MRI is a quantitative imaging modality by nature, the sets of streamlines reconstructed by tractography are not truly quantitative (Jones and Cercignani, 2010; Jbabdi and Johansen-Berg, 2011; Jones et al., 2013). Microstructure-informed tractography (Sherbondy et al., 2009, 2010; Smith et al., 2013, 2015; Pestilli et al., 2014; Daducci et al., 2015b, 2016; Girard et al., 2017) is a recent methodological advance which aims to overcome such limitations by complementing tractography with biophysical models of the tissue microstructure. One of the recent proposed methods is the *Convex Optimization Modeling for Microstructure Informed Tractography* (COMMIT) (Daducci et al., 2015b). COMMIT assigns contibutions to the signal to each reconstructed streamline according to a microstructural forward-model and attempts to express all the acquired DW-MRI signals as a linear combination of the contributions arising from the whole set of streamlines:

$$\mathbf{y} = \mathbf{A}\boldsymbol{x} + \eta, \tag{1}$$

where $\boldsymbol{y}$ contains the DW-MRI measurements in all voxels of the white matter, $\mathbf{A}$ is a matrix that accounts for the signal contributions of the streamlines in each voxel according to a given multi-compartment model (Panagiotaki et al., 2012) (possibly in addition to local voxel-wise contributions of tissue compartments, e.g., cerebrospinal fluid) and η is the acquisition noise. The unknown contributions $\boldsymbol{x}$ of all the compartments can then be efficiently estimated by solving the inverse problem using non-negative least squares:

$$\underset{\boldsymbol{x}\geq 0}{\operatorname{argmin}} \; ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}||_2^2. \tag{2}$$

Similarly to other filtering approaches, COMMIT assumes that the contributions of the streamlines are constant along their trajectories. More information on the method can be found in the original COMMIT manuscript (Daducci et al., 2015b).

3. MATERIALS AND METHODS

3.1. Bundle-Specific Estimation

To enable estimation of the axon diameter index of individual bundles, similarly to the recently proposed COMMIT-T_2 method (Barakovic et al., 2021), we extended the COMMIT framework with the Cylinder-Zeppelin-Ball model (Panagiotaki et al., 2012). The new formulation, COMMIT$_{\text{AxSize}}$, is presented in **Figure 2**. The proposed method *considers each streamline as consisting of a population of axons with an unknown distribution of diameters*, which must be estimated. The forward model (columns of the matrix $\mathbf{A}$) was the DW-MRI signal arising from axons represented as parallel cylinders oriented in parallel to the tangent to the streamline in the voxel and with fixed diameters and fixed longitudinal diffusivity $d_{\parallel}$. To account for different contributions arising from axons with distinct diameters, we considered 12 columns for each streamline corresponding to 12 cylinders with equally-spaced diameters in the range 1.5 µm to 7 µm. We modeled the extra-axonal compartment with anisotropic tensors, i.e., Zeppelins, having the same longitudinal diffusivity $d_{\parallel}$. To capture different geometries of the extra-axonal space in every voxel, we considered multiple Zeppelins in every voxel, each with a distinct perpendicular diffusivity $d_{\perp}$. Moreover, a distinct set of Zeppelins was included in $\mathbf{A}$ for every principal diffusion direction in a voxel. Finally, the cerebrospinal fluid was modeled as an isotropic tensor, i.e., Ball, with fixed diffusivity d_{iso}; an independent contribution was assigned to each voxel. The physical parameters were set according to values found in the literature (Alexander et al., 2010; Zhang et al., 2011b, 2012; Le Bihan and Iima, 2015): $d_{\parallel}$=1.7 × 10^{-3} mm^2 s^{-1}, d_{iso}=3.0 × 10^{-3} mm^2 s^{-1}, and four reasonable values equally-spaced from 0.5 × 10^{-3} mm^2 s^{-1} to 1.0 × 10^{-3} mm^2 s^{-1} for $d_{\perp}$.

The *axon diameter index of a streamline*, can be estimated from the coefficients $\mathbf{x}$ computed by COMMIT$_{\text{AxSize}}$ as done in the ActiveAx$_{\text{AMICO}}$ method (Daducci et al., 2015a); in fact, the 12 contributions corresponding to a given streamline represent its *volume-weighted cylinder diameter distribution*. Unlike in Assaf et al. (2008), no assumptions are made on the axon diameter distribution to be estimated. The cylinder diameter distribution can be defined for a bundle, i.e., a group of streamlines coursing through a specific region of interest (ROI). Hence, we grouped

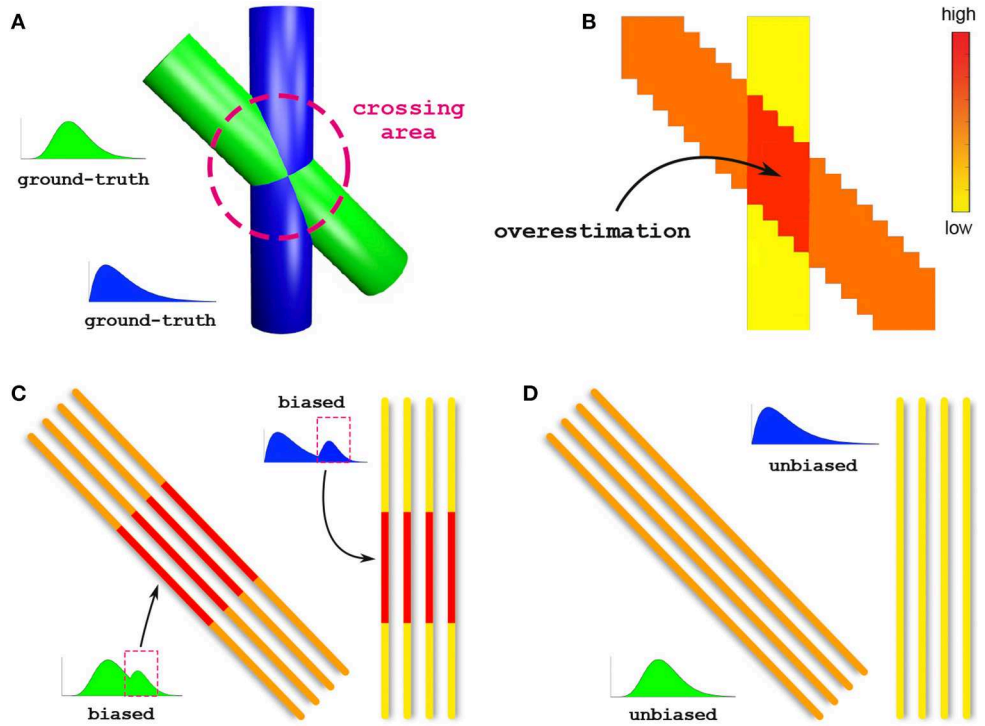


FIGURE 1 | *Voxel-wise vs. bundle-specific axon diameter estimation.* **(A)** Schematic illustration of two crossing fiber populations characterized by different compositions: the green bundle contains larger axons than the blue. **(B)** Axon diameter estimation using a voxel-wise approach; for simplicity, we report the estimated mean diameters. The arrow points to the crossing region where such methods are known to especially suffer from overestimation. **(C)** Characterization of the axon diameter of a white matter bundle is typically done by averaging, along its entire course, the values previously estimated in every voxel; this indirect procedure is affected by such overestimated voxels and leads to biased results. **(D)** Estimation of bundle-specific axon diameter.

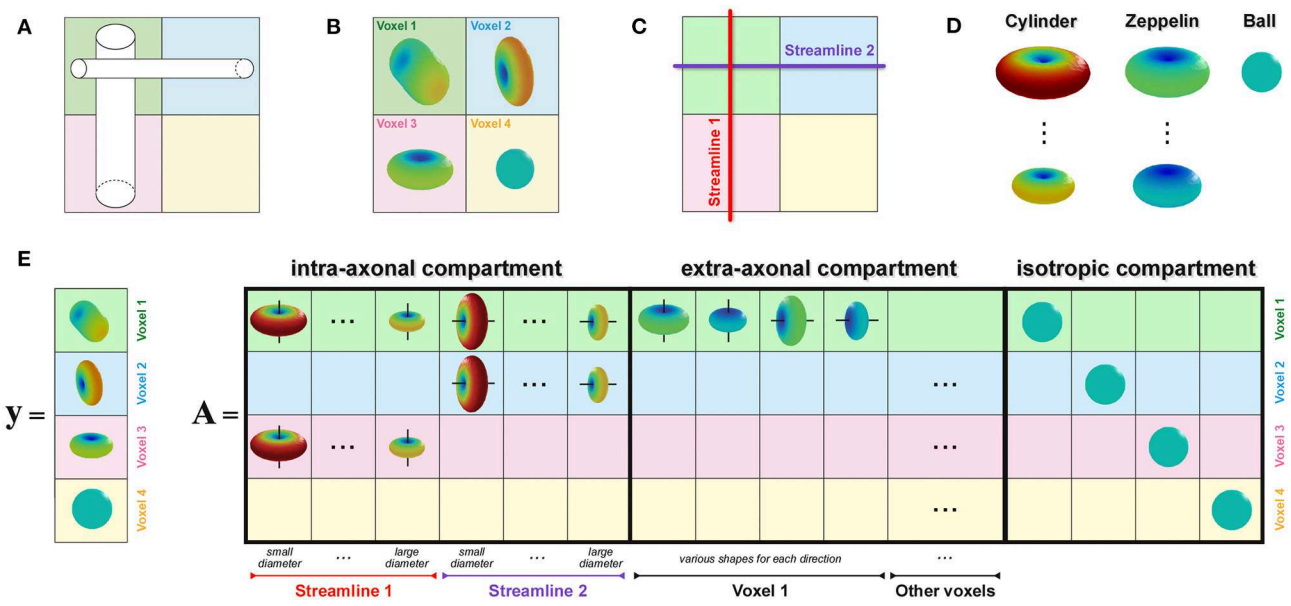


FIGURE 2 | *How to enable estimation of bundle-specific axon diameter index.* **(A)** Simple crossing configuration of two fiber populations with different axon compositions, i.e., the vertical one is composed of larger axons than the horizontal, to illustrate the construction of the proposed formulation. **(B)** Corresponding DW-MRI signal in four representative voxels. **(C)** Example of two possible streamlines reconstructed with tractography. **(D)** Visual representation of the response functions in the Cylinder-Zeppelin-Ball forward model for each compartment. **(E)** The vector $\boldsymbol{y}$ contains a concatenation of the DW-MRI signal acquired in all voxels, while the matrix $\boldsymbol{A}$ is constructed by combining the response functions with the local orientations of the streamlines in each voxel.

TABLE 1 | DW-MRI acquisition protocol parameters.

b-value (s mm^{-2})	δ (ms)	Δ (ms)	G (mT m^{-1})	directions
1,000	7	17.3	138	30
4,000	7	17.3	276	60
1,000	7	30	102	30
4,000	7	30	203	60
1,000	7	42	85	30
4,000	7	42	169	60
1,000	7	55	74	30
4,000	7	55	175	60

The images were acquired using a 2 mm isotropic resolution and a matrix size of 110 × 110. The echo-time (TE) was 80 ms and the repetition time (TR) was 3900 ms.

streamlines sharing the same anatomical pathways in bundles as defined by an anatomical atlas. We then calculated the axon diameter index of a bundle by performing the weighted sum, column by column, of the cylinder signature of all streamlines of the bundle.

To facilitate visual inspection of the results, we extended the Axon Diameter Index (ADI) (Alexander et al., 2010) to streamlines (sADI), which is the mean of the distribution, and colored all streamlines accordingly. To compute the sADI, we excluded the contributions of the smallest (1.5 μm) and the biggest (7 μm) cylinder diameters. This is for two reasons: i) the used DW-MRI acquisition was shown to be insensitive to diameters smaller than 2 μm (Nilsson et al., 2017). ii) We found that the smallest cylinder captures, only partially, the signal of axons from 0 μm to 1.5 μm, and the biggest cylinder captures the signal of axons above 7 μm; hence, the coefficients corresponding to those columns of $\boldsymbol{A}$ are unreliable for the computation of the sADI. Simulations were performed to validate this assumption, see **Supplementary Materials**.

3.2. Data Acquisition

3.2.1. *In-vivo* Human Data

In vivo human data were acquired from 3 healthy volunteers on a Siemens Connectom 3 T MRI system (Cardiff University Brain Research Centre, Cardiff, Wales). The studies involving human participants were reviewed and approved by The School of Psychology Ethics Committee, Cardiff University. All participants provided written informed consent to participate in this study. Each subject was imaged five times over 2 weeks using the same DW-MRI acquisition protocol. The DW-MRI acquisition protocol used is the following: echo-time (TE) 80 ms, repetition time (TR) 3.900 ms, matrix size 110 × 110, 2 mm isotropic resolution. Other protocol parameters are reported in the **Table 1**.

Five non-diffusion weighted images (b0) were acquired, including one in reverse phase encoding. A 1 mm isotropic resolution T1-weighted anatomical image was also acquired, using a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence: TE = 2 ms, inversion time = 857 ms, TR = 2.300 ms, matrix size = 256 × 256, flip angle = 9°.

3.2.2. Simulation Data

A numerical phantom was generated with a 45° crossing configuration between two bundles, from which, main directions were obtained at each voxel. The intra-axonal and extra-axonal signals were generated separately and then merged to generate unique numerical phantom (Rensonnet et al., 2018). For each bundle, the DW-MRI intra-axonal signal was simulated, using a distribution of parallel cylinders (Van Gelderen et al., 1994) following a gamma distribution. The first bundle had a gamma distribution with volume weighted mean diameter of 2.70 (shape = 3.2734 and scale = 0.2556). For the second bundle, the volume weighted mean diameter was 4.00 (shape = 3.5027 and scale = 0.3655). The extra-axonal signal was generated using a tensor with perpendicular diffusivity adapted to the local intra-axonal volume fraction, following the tortuosity approximation (Szafer et al., 1995). In single fiber voxels, the intra-axonal signal fractions were set to 0.3 and 0.6 for the vertical and diagonal bundles, respectively (i.e., the extra-axonal signal was a tensor with perpendicular diffusivity equal to $0.7 \times D$ and $0.4 \times D$, respectively). The crossing voxels had an intra-axonal volume fraction of 0.9 (i.e., the extra-axonal signals were generated with a perpendicular diffusivity equal to $0.1 \times D$). All signals were summed to have a total signal fraction of 1 in each voxel. The diffusivity of the simulations were fixed to $D = 1.7 \times 10^{-3}$ mm^2 s^{-1} (Alexander et al., 2010; Zhang et al., 2011b, 2012), both for intra-axonal and extra-axonal signals. The resulting dataset was corrupted with various levels of Rician noise. Furthermore, four additional dataset were generated, adding voxel-wise dispersion using a Watson distribution with k = 4, 8, 12, 16 (Zhang et al., 2011b, 2012).

For the voxel-wise estimation, the ADI for each voxel was estimated with the ActiveAx method (Alexander et al., 2010) implemented in the AMICO framework (Daducci et al., 2015b). For the COMMIT$_{\text{AxSize}}$ method, the bundle-specific axon diameter index were estimated using both the ground-truth bundle trajectories and using the MrTrix3 second-order integration over Fiber Orientation Distribution (iFOD2) algorithm generating approximately 1,000 streamlines per bundle. Streamlines not ending at the bundle extremities were removed before processing with COMMIT$_{\text{AxSize}}$.

3.3. Data Pre-processing

The anatomical T1-weighted image was registered to the preprocessed average b0 image using FSL/FLIRT (Jenkinson and Smith, 2001) using rigid-body registration. The white matter and gray matter masks were estimated using FSL/FAST (Zhang et al., 2001). The brain cortical parcellation was performed using FreeSurfer (Destrieux et al., 2010). The DW-MRI images were corrected for magnetic field inhomogeneities, eddy currents (Andersson and Sotiropoulos, 2016) and motion using the TOPUP (Graham et al., 2017), and EDDY tools of FSL (Jenkinson and Smith, 2001). Subsequently, gradient non-linearity correction was performed (Jovicich et al., 2006; Rudrapatna et al., 2021). The shell with diffusion time Δ = 17.3 ms, G = 276 mT m^{-1} and b-value = 4000 s mm^{-2} was used to perform Constrained Spherical Deconvolution (CSD) (Tournier et al., 2007). Tractography was then performed using

iFOD2 algorithm (Tournier et al., 2012), generating 10,000,000 streamlines seeding from the white matter mask. Streamlines not reaching the gray matter mask were removed. To make the computational time practical, a sub-set of 300,000 streamlines was randomly selected for each DW-MRI dataset.

3.4. Analysis of Specific Neuronal Connections

We report the sADI estimated for the streamlines passing through individual sectors of corpus callosum (CC) and of the posterior limb of the internal capsule (PIC). The midsagittal section of the CC was outlined using to the FreeSurfer parcellation, and the transverse section of the PIC was manually outlined on the T1-weighted image by an expert anatomist. The skeletons (longitudinal centerline) of both regions were computed and then subdivided into equally-spaced segments. The boundaries of each sector were drawn roughly perpendicular to the skeleton by associating all voxels within the outlines to their closest segment. We fixed 11 regions of interest (ROIs) in the CC showed in **Figure 3A**, and 6 ROIs in the PIC showed for each hemisphere in **Figure 3E**.

Our *in vivo* study is focused on two well-characterized axonal tracts: the CC and the PIC. The CC has been well studied in the past with different methodologies, including DW-MRI (Barazany et al., 2009; Alexander et al., 2010). The PIC has been less studied with DW-MRI but is extremely important since it is traversed by cortico-descending axons involved in motor control, whose lesions lead to irreversible paralysis. Moreover, we concentrate the analysis on these two bundles since they are known to have a sufficiently large axon diameter, **Figure 6**. To study the topology of bundles, the CC and the PIC were segmented and subdivided in, respectively, 11 and 6 equal ROIs normalized for different individuals as described in section 3.4. The streamlines passing through regions of interest (ROIs) corresponding to these sectors were selected, and we analyzed their projections to and from the cortex. These projections correspond to corticofugal and corticopetal (for the CC) connections since DW-MRI does not distinguish the direction of the connections. Bundles of streamlines systematically organized from anterior to posterior connect the CC to similarly ordered slabs of cortex extending from the cingulate gyrus to the lateral sulcus (see **Figure 3**). This is usually neglected the aspect of CC topology, albeit already shown by tracer injections in the CC of the cat (Nakamura and Kanaseki, 1989), and is compatible with the ordering of CC connections already described with DW-MRI (Hofer et al., 2015). Also, anteroposteriorly organized bundles of streamlines connect the sectors of PIC to anteroposterior cortical territories, compatible with the topology shown by tracer injections and DW-MRI in monkeys (Morecraft et al., 2017) and DW-MRI in humans (Archer et al., 2018).

3.5. Comparison With Histology

The fiber composition of the CC obtained with $\text{COMMIT}_{\text{AxSize}}$ was compared with postmortem measurements from a previous study (Caminiti et al., 2009); however, to evaluate the impact of histological sampling one of the sectors was measured again (see **Supplementary Figure 8**). Between 451 and 1934 axons stained for myelin were measured in CC sectors crossed by axons connecting the prefrontal, motor, parietal and visual cortices. From the histological data, we estimated the histogram of diameters in each sector. However, since DW-MRI estimates the signal fractions that are related to the volume occupied by axons of different diameter, not their number, the data was converted to volume-weighted distributions, to allow comparison with the DW-MRI estimates. In the absence of human data, the *in vivo* estimates of the PIC were compared with measurements of axons stained for myelin in the monkey PIC (Innocenti et al., 2018).

4. RESULTS AND DISCUSSION

4.1. Numerical Simulations

Figure 4 compares the estimated ADI obtained using the conventional voxel-wise procedure and the proposed bundle-specific $\text{COMMIT}_{\text{AxSize}}$ on the numerical phantom described in Section 3.2. The results from $\text{COMMIT}_{\text{AxSize}}$ show more consistent estimates of the bundles mean cylinder diameter, compared to the voxel-wise method, in particular at low SNR (**Figure 4** first column). Moreover, at SNR $=$ 50 $\text{COMMIT}_{\text{AxSize}}$ estimated on average a sADI of 2.90 μm and 4.01 μm compare to 3.37 μm and 4.25 μm for the $\text{ActiveAx}_{\text{AMICO}}$ (the mean ground-truth diameter of each bundle is 2.7 μm and 4.0 μm, respectively). The estimates provided by $\text{COMMIT}_{\text{AxSize}}$ are both more robust to noise, and closer to the real values, when compared to voxel-wise estimates.

The second columns of **Figure 4** show the ADI estimates at SNR $=$ 50, changing the dictionary diameter sampling of both methods. Results improve for $\text{COMMIT}_{\text{AxSize}}$ increasing the sampling density, in particular for the bundle with the smallest mean diameter. Results of the voxel-wise method are similar using all dictionaries, with the best performances using 12 or 15 values. In all cases, $\text{COMMIT}_{\text{AxSize}}$ outperformed the voxel-wise method. Although more elements in the dictionary improve the estimation, the optimisation problem becomes harder and increases the computation requirements. Nonetheless, a dictionary sampling of 12 columns in matrix A provides a reasonable estimate on synthetic data, while keeping the computation requirement feasible for *in vivo* data.

Finally, **Figure 4** (third column) show the ADI estimated in the same phantom, but including various levels of dispersion (SNR $=$ 50, dictionary diameter sampling of 12). Rather than using the ground-truth cylinder trajectories, we used probabilistic tractography to estimate their trajectories, capturing the dispersion information from the data. The rightmost boxplot shows the estimates using the probabilistic tractography with no dispersion ($\kappa = \text{inf}$). Using the probabilistic tractography streamlines, $\text{COMMIT}_{\text{AxSize}}$ shows an underestimation of the mean diameter when compared to the ground-truth bundle trajectories. However, the increase in dispersion (lower κ value) show a systematic over-estimation of the mean diameter of the largest bundles, and little effect on the bundle with the smallest cylinder diameter. However, the trend changes at $\kappa = 4$, where both bundle ADI are estimated between 2.9 μm and 3.5 μm. This could be explained by the inability of probabilistic

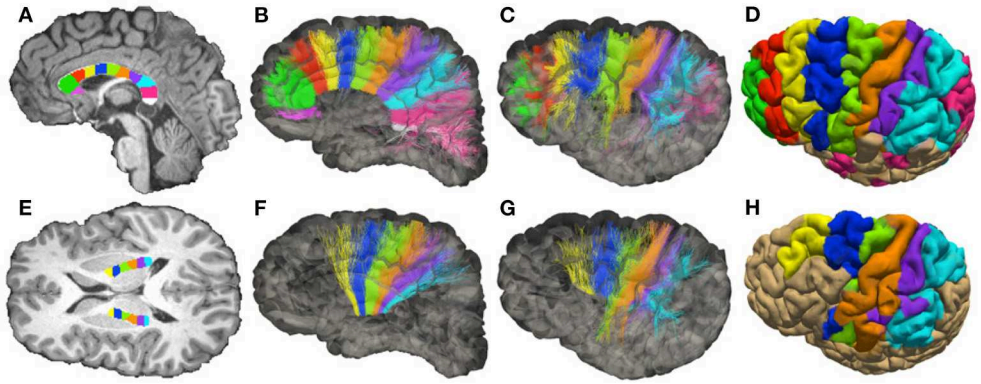


FIGURE 3 | Topology of fibers in the Corpus Callosum (CC) and posterior limb of the internal capsule (PIC), reconstructed with DW-MRI tractography of a single healthy volunteer. **(A)** Subdivision of the mid-sagittal section of the CC in 11 sectors (corresponding to ROIs), see **Supplementary Figure 9**. **(B,C)** Streamlines colored according to the corresponding ROIs (medial and lateral views of the hemisphere). **(D)** Projection of the streamlines onto the pial surface. **(E)** Subdivision of PIC in 6 sectors (ROIs). **(F,G)** Streamlines colored according to the corresponding ROIs (medial and lateral view of the hemisphere). **(H)** Projection of the streamlines onto the pial surface.

tractography to properly capture this high level of dispersion. Although $\text{COMMIT}_{\text{AxSize}}$ cannot fully model the dispersion, the estimate along the streamlines provides more robust estimates of the bundle diameter than the voxel-wise method.

Moreover, contrary to the voxel-wise method, $\text{COMMIT}_{\text{AxSize}}$ can disentangle bundles in crossing configurations and provide a reliable bundle-specific ADI in those areas. Something not achievable robustly with a voxel-wise method assuming a single fiber population. These numerical experiments showed the benefit of $\text{COMMIT}_{\text{AxSize}}$ when estimating axon diameter indexes.

4.2. *In vivo* Data

Figure 5 shows the streamlines passing through the CC (A) and the PIC (C), colored following their corresponding sADI. **Figures 5B,D** show sADI projected onto the pial surface. In both bundles we studied, the largest sADI were found in sectors of PIC traversed by axons connecting the motor cortex (BA 4) while smaller sADI were found for other areas. This visualization reveals that streamlines with larger sADI connect the CC to the precentral gyrus, corresponding to the primary motor cortex (M1; Broadman area BA 4), the more lateral part of premotor cortex (BA 6), and the postcentral gyrus (BA 3,1,2) corresponding to the primary somatosensory cortex (S1). Streamlines with progressively smaller sADI terminate in the medial premotor cortex (BA 6) and the parietal cortex (BA 5,7 and 40) and still smaller sADI in the rostral prefrontal cortex (BA 8 and 9) and BA 44 and 45. In case of the CC, human postmortem material was used to validate the estimates obtained with our novel technique. The comparison was performed in four different ROIs. **Figure 6** shows that the bundle sADI estimated with $\text{COMMIT}_{\text{AxSize}}$ closely corresponds to the histological estimates within the DW-MRI range of sensitivity.

In monkey species (Caminiti et al., 2009; Tomasi et al., 2012), a hierarchy of axon diameters exists with thicker and faster-conducting axons connecting the motor and somatosensory cortices, thinner and slower axons elsewhere. The streamlines coursing in the PIC were color-coded as above according to their estimated sADI. These with the high sADI mapped onto the dorsal part of the precentral BA 4 (M1) and postcentral (BA 3,1,2; S1) gyrus. Progressively smaller sADI mapped onto the parietal cortex (BA 5 and 7) and the premotor cortex (BA 6) and still smaller sADI onto the rostral prefrontal cortex (BA 8 and 9). This arrangement is similar to that demonstrated with injections of anterogradely transported tracers in corresponding areas of the monkey, although in the monkey the diameter of axons originating in the precentral gyrus exceeds that of axons originating in the postcentral gyrus (Innocenti et al., 2018). Identical findings were reproduced for different sectors of the CC and PIC in three subjects and five times for each subject, **Figure 6** and **Supplementary Figure 10**.

Current technologies restrict the resolution of axon diameters to about 2.0 μm (Nilsson et al., 2017). Nevertheless, axons with larger diameter show a detectable contrast according to our simulations, see **Supplementary Materials**. Despite this limitation, since large axons are found preferentially in specific pathways, their absence in the expected pathways, or abnormal presence in unexpected pathways can disclose the neural basis of specific neurological or psychiatric pathologies (DeLuca et al., 2004; Zikopoulos and Barbas, 2013; Huang et al., 2016; Judson et al., 2017; Golden et al., 2020) and, possibly, of individual skills (de Manzano and Ullén, 2018).

4.3. Strengths and Limitations

Our proposed bundle-specific approach allows investigation of the intrinsic axon composition of white matter pathways as opposed to sampling their composition at discrete locations along their course, making no assumptions on the axon composition of a bundle. As the DW-MRI signal in each voxel is expressed as the combined contributions of multiple intersecting streamlines, our method naturally handles the presence of different pathways within a voxel and allows their individual

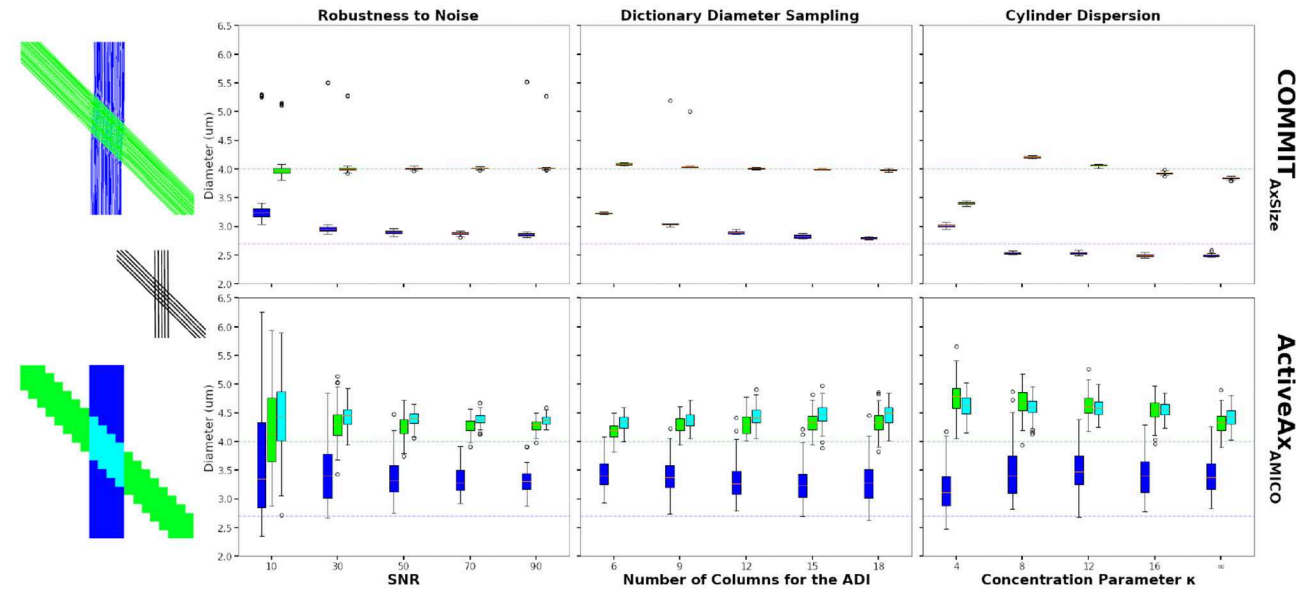


FIGURE 4 | Axon diameter indexes (ADI) estimated on the 45-degrees crossing synthetic phantom. The figures show the bundle-specific ADI estimated using the $\text{COMMIT}_{\text{AxSize}}$ method (top row) and the voxel-wise ADI estimated using the $\text{ActiveAx}_{\text{AMICO}}$ method (bottom row). The boxplots show the results 50 different Rician noise realisations for $\text{COMMIT}_{\text{AxSize}}$ and 100 voxels sampled in each single bundle areas and in the crossing area, for the $\text{ActiveAx}_{\text{AMICO}}$ method. The mean ground truth cylinder diameter of the green and blue bundles are 4.0 μm, and 2.7 μm, respectively (colored dashed line). The first boxplot column show the estimated ADI at various SNR, the second column using various number of columns in matrix A to compute the ADI and the last column shows the estimates for data with various levels of dispersion.

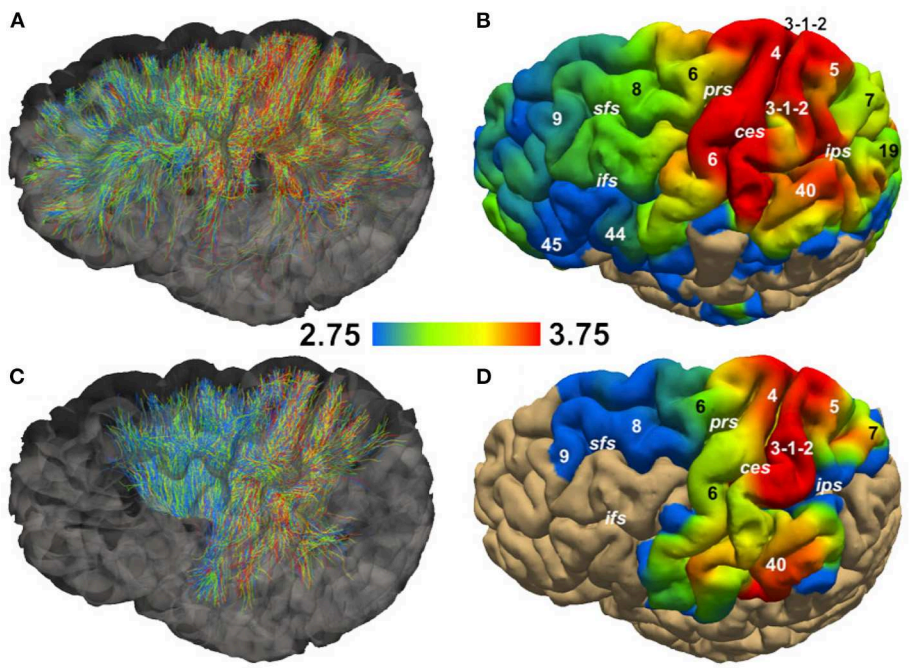


FIGURE 5 | Streamline Axon Diameter Index (sADI) in the Corpus Callosum (CC) and posterior limb of the internal capsule (PIC) of a single healthy volunteer. **(A,C)** show streamlines colored according to their sADI. **(B,D)** show the projection of streamlines'sADI onto the pial surface; colors correspond to the sADI averaged across streamlines. Abbreviations: ces, central sulcus; ifs, inferior frontal sulcus; ips, interparietal sulcus; prs, precentral sulcus; sfs, superior frontal sulcus. Numbers correspond to Brodman areas.

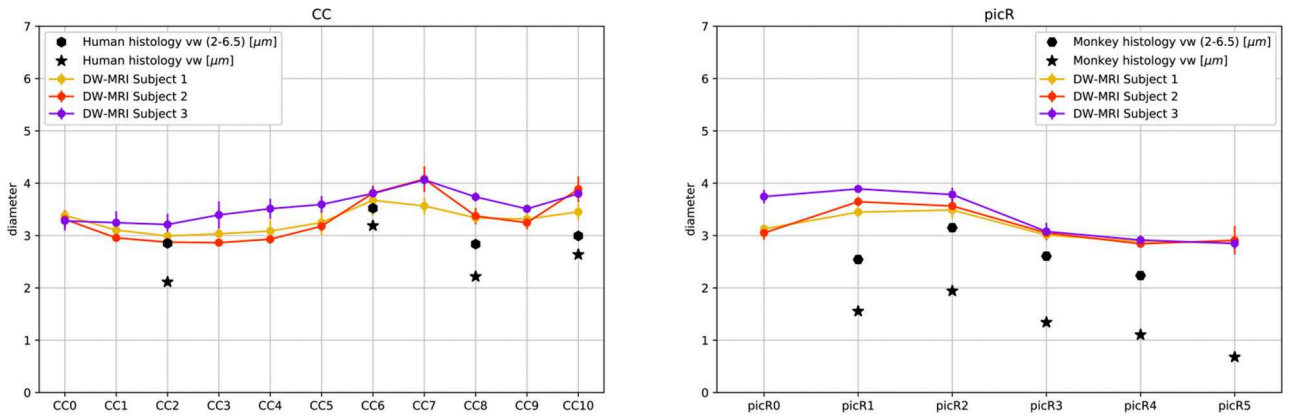


FIGURE 6 | DW-MRI cylinder diameter distribution of CC and PIC sectors compared to histological mean volume-weighted axon diameter from human (CC) and monkey (PIC) histology. Star markers represent the means of volume weighted histological distributions; hexagon markers represent the means of volume weighted cylinder distributions in the range 2.0μm to 6.5μm.

contributions to be decoupled. This contrasts with methods in which a 'powder average' of the diffusion-weighted signal is taken as part of the axon diameter estimation process (Veraart et al., 2018). On the contrary, in voxels with such complex fiber configurations, the current voxel-wise estimation approaches provide biased estimates. Another advantage of our approach is that the cylinder diameter distribution of a bundle could be mapped onto the cortex where it originates (and/or terminates), eliminating the ambiguities of following axon diameters at selected locations along the white matter pathways (Assaf et al., 2008; Barazany et al., 2009; Alexander et al., 2010).

One could argue that it has been demonstrated that although the axonal diameter of single axons can undergo local changes along its trajectory (Lee et al., 2019); the axon diameter distribution of diameters in a pathway remains stable over long distances (Tomasi et al., 2012).

Nevertheless, we stress that a streamline represents a group of axons that share a similar trajectory; thus, our method estimates an average diameter for the represented group. Moreover, by discretizing the intra-axonal signal in the contributions arising from multiple impermeable cylinders (Van Gelderen et al., 1994), each streamline can be composed of a different amount of cylinders with different diameters without imposing any prior on the eventual distribution.

By decomposing the signal of each voxel into three components (intra-axonal, extra-axonal and isotropic compartments) and regularizing the intra-axonal signal fractions along streamlines, we were able to detect the signal fractions corresponding to each component. In particular, we discretized the signal coming from each cylinder diameter using the formula for impermeable cylinders of Van Gelderen et al. (1994) and what we estimated through the $\text{COMMIT}_{\text{AxSize}}$ method was the weighting factor in front of each diameter d_i, which is of the order of d_i^2 (Burcaw et al., 2015). Similarly, we discretized the signal coming from the extra-axonal compartment in two main components, parallel and perpendicular, along each principal direction. Both components were fixed using physically plausible constant values for the diffusion coefficient (Alexander et al., 2010; Zhang et al., 2012). Moreover, for the perpendicular direction, we accounted for four possible diffusion coefficients (i.e., in each voxel for each main direction we estimated five possible fractions of extra-axonal signal: one parallel to the fiber population and four perpendiculars to it). The remaining fraction of the signal in each voxel was then captured by the signal contribution of an isotropic compartment with fixed diffusivity. With these parameters, we do not account for the eventual residual time dependence of the extra-axonal diffusion tensor. However, by allowing a different fraction for each discretized value in the perpendicular direction, we account for a positive contribution of this compartment. Indeed, although we acquired data with strong gradients, we are not in a regime for which the extra-axonal signal may not be completely suppressed (Veraart et al., 2019). The model we used in the signal discretization, as well as the acquisition parameters chosen for the DW-MRI sequence, can be improved to be more sensitive to microstructure features.

Recent studies have suggested that axons may vary in diameter along their length, may present undulations and microscopic orientation dispersion (Brabec et al., 2020; Lee et al., 2020a; Rafael-Patino et al., 2020) which impacts the estimate obtained with DW-MRI (Lee et al., 2020b). The impact of those tissue properties on $\text{COMMIT}_{\text{AxSize}}$ will be addressed in future studies. Furthermore, the volume fractions selected for our simulation experiment were limited by the choice of substrate simulator (Rafael-Patino et al., 2020). Future numerical simulations will address more complex configurations, and more realistic substrates will be constructed (e.g., varying volume fractions and diameter distributions).

In this study, we proposed to estimate the bundle-specific axon diameter index implementing the Cylinder-Zeppelin-Ball forward model. We believe that the selection of the optimal microstructure forward model and dictionary parameters

could be improved in future studies. In particular, the dictionary discretization used (i.e., 12 values for the intra-axonal component) and fixed values for diffusivities, which may cause loss of accuracy (Jelescu et al., 2016), will be two aspects to explore extensively. Moreover, various tractography algorithms will be tested to build the $COMMIT_{AxSize}$ dictionary (e.g., deterministic and probabilistic, and tractography parameters). Future works will address these aspects, including the exploiting different diffusion acquisition protocols, using similar approaches as in Drobnjak et al. (2016) and Nilsson et al. (2017), to find an optimal set of parameters and protocols improving the sentitivity to the tissue properties (Lampinen et al., 2017). Another important aspect to mention is that the $COMMIT_{AxSize}$ may be inaccurate in the diseased brain, affected by focal lesions along white matter tracts. However, it may be applicable to developmental disorders, psychiatric disorders and neurological diseases such as epilepsy.

5. CONCLUSION

In this paper, we focused on the non-invasive characterization of the composition of central nervous system pathways in the living human brain from DW-MRI acquisitions. In particular, we tackled some fundamental limitations of current voxel-wise techniques and proposed a novel formulation to estimate the axon diameter index of a fiber bundle all along its trajectory, rather than sampling it at a few selective locations along its course. We compared our bundle-specific approach to the state-of-the-art voxel-wise methods, both on synthetic and *in vivo* human brain data, comparing our findings with histological measurements in two well-studied cortical pathways. Our results demonstrated the feasibility and the benefits of our proposed formulation. Moreover, the bundle composition estimated agree with histology and known anatomy. Further studies could extend the present approach to other pathways in the central nervous system, enhancing the human connectome enterprise (Craddock et al., 2013; Jbabdi et al., 2015; Glasser et al., 2016).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

MB, GG, DR, SS, MD, CG, DJ, GI, J-PT, and AD conceptualized the problem. MB and AD developed, implemented, and tested the technical framework. DJ provided the MRI data. MB and GI validated the framework. GI contextualized the framework with a biological and application-oriented perspective. MB, GG, SS, GI, and AD wrote the manuscript.

FUNDING

The work was supported by the Swiss National Science Foundation 31003A_157063 and 205320_175974 as well as by the Rita Levi Montalcini Programme for young researchers of the Italian Ministry of Education, University and Research (MIUR). The MRI data were acquired at the UK National Facility for *in vivo* MR Imaging of Human Tissue Microstructure funded by the EPSRC (grant EP/M029778/1), and The Wolfson Foundation. DJ is supported by a Wellcome Trust Investigator Award (096646/Z/11/Z) and a Wellcome Trust Strategic Award (104943/Z/14/Z).

ACKNOWLEDGMENTS

We want to acknowledge U. Rudrapatna, C. M.W. Tax and the CUBRIC center for computational resources and for the acquisition of the DW-MRI data; R.A.W. Galuske and the Technische Universitt Darmstadt for human post-mortem histological samples. We also acknowledge E.J. Canales-Rodriguez for fruitful discussions about optimal estimation of parameters. We acknowledge access to the facilities and expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Ecole polytechnique fdrale de Lausanne (EPFL), University of Geneva (UNIGE) and Geneva University Hospitals (HUG).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.646034/full#supplementary-material

REFERENCES

Aboitiz, F., Scheibel, A. B., Fisher, R. S., and Zaidel, E. (1992). Fiber composition of the human corpus callosum. *Brain Res.* 598, 143–153. doi: 10.1016/0006-8993(92)90178-c

Alexander, D. C., Hubbard, P. L., Hall, M. G., Moore, E. A., Ptito, M., Parker, G. J., et al. (2010). Orientationally invariant indices of axon diameter and density from diffusion MRI. *Neuroimage* 52, 1374–1389. doi: 10.1016/j.neuroimage.2010.05.043

Andersson, J. L., and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* 125, 1063–1078.

Archer, D. B., Vaillancourt, D. E., and Coombes, S. A. (2018). A template and probabilistic atlas of the human sensorimotor tracts using diffusion MRI. *Cereb. Cortex* 28, 1685–1699. doi: 10.1093/cercor/bhx066

Assaf, Y., Blumenfeld-Katzir, T., Yovel, Y., and Basser, P. J. (2008). AxCaliber: a method for measuring axon diameter distribution from diffusion MRI. *Magn. Reson. Med*. 59, 1347–1354. doi: 10.1002/mrm.21577

Auria, A., Romascano, D., Canales-Rodriguez, E., Wiaux, Y., Dirby, T. B., Alexander, D., et al. (2015). "Accelerated microstructure imaging via convex optimisation for regions with multiple fibres (AMICOx)," in *2015 IEEE International Conference on Image Processing (ICIP)* (Quebec).

Barakovic, M., Tax, C. M., Rudrapatna, U., Chamberland, M., Rafael-Patino, J., Granziera, C., et al. (2021). Resolving bundle-specific intra-axonal t2 values within a voxel using diffusion-relaxation tract-based estimation. *Neuroimage* 227:117617.

Barazany, D., Basser, P. J., and Assaf, Y. (2009). *in vivo* measurement of axon diameter distribution in the corpus callosum of rat brain. *Brain* 132(Pt 5), 1210–1220. doi: 10.1093/brain/awp042

Barazany, D., Jones, D., and Assaf, Y. (2011). "AxCaliber 3D," in *In Proceedings of the 19th Scientific Meeting of ISMRM* (Montreal, QC).

Beaulieu, C., and Allen, P. (1994). Water diffusion in the giant axon of the squid: implications for diffusion-weighted MRI of the nervous system. *Magn. Reson. Med*. 32, 579–583.

Brabec, J., Lasi, Ċ, S., and Nilsson, M. (2020). Time-dependent diffusion in undulating thin fibers: Impact on axon diameter estimation. *NMR Biomed*. 33:e4187. doi: 10.1002/nbm.4187

Burcaw, L., Fieremans, E., and Novikov, D. (2015). Mesoscopic structure of neuronal tracts from time-dependent diffusion. *Neuroimage* 114, 18–37.

Caminiti, R., Ghaziri, H., Galuske, R., Hof, P. R., and Innocenti, G. M. (2009). Evolution amplified processing with temporally dispersed slow neuronal connectivity in primates. *Proc. Natl. Acad. Sci. U.S.A*. 106, 19551–19556.

Craddock, R. C., Jbabdi, S., Yan, C. G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* 10, 524–539. doi: 10.1038/nmeth.2482

Daducci, A., Canales-Rodriguez, E., Zhang, H., Dyrby, T., Alexander, D., and Thiran, J.-P. (2015a). Accelerated microstructure imaging via convex optimization (AMICO) from diffusion MRI data. *Neuroimage* 105, 32–44.

Daducci, A., Dal Palù, A., Lemkaddem, A., and Thiran, J. P. (2015b). COMMIT: convex optimization modeling for microstructure informed tractography. *IEEE Trans. Med. imaging* 34, 246–257. doi: 10.1109/TMI.2014.2352414

Daducci, A., Palú, A. D., Descoteaux, M., and Thiran, J. P. (2016). Microstructure informed tractography: pitfalls and open challenges. *Front. Neurosci*. 10:247. doi: 10.3389/fnins.2016.00247

de Manzano, Ö. and Ullén, F. (2018). Same Genes, different brains: neuroanatomical differences between monozygotic twins discordant for musical training. *Cereb Cortex* 28, 387–394. doi: 10.1093/cercor/bhx299

DeLuca, G. C., Ebers, G. C., and Esiri, M. M. (2004). Axonal loss in multiple sclerosis: a pathological survey of the corticospinal and sensory tracts. *Brain* 127(Pt 5), 1009–1018. doi: 10.1093/brain/awh118

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010

Drakesmith, M., Harms, R., Rudrapatna, S. U., Parker, G. D., Evans, C. J., and Jones, D. K. (2019). Estimating axon conduction velocity *in vivo* from microstructural MRI. *Neuroimage* 203:116186. doi: 10.1016/j.neuroimage.2019.116186

Drobnjak, I., Siow, B., and Alexander, D. C. (2010). Optimizing gradient waveforms for microstructure sensitivity in diffusion-weighted MR. *J. Magn. Reson.* 206, 41–51. doi: 10.1016/j.jmr.2010.05.017

Drobnjak, I., Zhang, H., Ianu, A., Kaden, E., and Alexander, D. C. (2016). Pgse, ogse, and sensitivity to axon diameter in diffusion mri: Insight from a simulation study. *Magn. Reson. Med.* 75, 688–700. doi: 10.1002/mrm.25631

Dyrby, T., Sogaard, L., Hall, M., Ptito, M., and Alexander, D. (2013). Contrast and stability of the axon diameter index from microstructure imaging with diffusion MRI. *Magn. Reson. Med*. 70, 711–721. doi: 10.1002/mrm.24501

Fan, Q., Nummenmaa, A., Witzel, T., Ohringer, N., Tian, Q., Setsompop, K., et al. (2020). Axon diameter index estimation independent of fiber orientation distribution using high-gradient diffusion MRI. *Neuroimage* 222:117197. doi: 10.1016/j.neuroimage.2020

Farooq, H., Xu, J., Nam, J., Keefe, D., Yacoub, E., Georgiou, T., et al. (2016). Microstructure Imaging of Crossing (MIX) White Matter Fibers from diffusion MRI. *Sci. Rep*. 6:38927. doi: 10.1038/srep38927

Girard, G., Daducci, A., Petit, L., Thiran, J.-P., Whittingstall, K., Deriche, R., et al. (2017). AxTract: toward microstructure informed tractography. *Hum. Brain Mapp*. 38, 5485–5500. doi: 10.1002/hbm.23741

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178. doi: 10.1038/nature18933

Golden, C. E. M., Yee, Y., Wang, V. X., Harony-Nicolas, H., Hof, P. R., Lerch, J. P., et al. (2020). Reduced axonal caliber and structural changes in a rat model of Fragile X syndrome with a deletion of a K-Homology domain of Fmr1. *Transl. Psychiatry* 10:280. doi: 10.1038/s41398-020-00943-x

Gore, J. C., Xu, J., Colvin, D. C., Yankeelov, T. E., Parsons, E. C., and Does, M. D. (2010). Characterization of tissue structure at varying length scales using temporal diffusion spectroscopy. *NMR Biomed.* 23, 745–756. doi: 10.1002/nbm.1531

Graham, M. S., Drobnjak, I., Jenkinson, M., and Zhang, H. (2017). Quantitative assessment of the susceptibility artefact and its interaction with motion in diffusion MRI. *PLoS ONE* 12:e0185647. doi: 10.1371/journal.pone.0185647

Harkins, K. D., Beaulieu, C., Xu, J., Gore, J. C., and Does, M. D. (2021). A simple estimate of axon size with diffusion MRI. *Neuroimage* 227:117619. doi: 10.1016/j.neuroimage.2020.117619

Hofer, S., Wang, X., Roeloffs, V., and Frahm, J. (2015). Single-shot T1 mapping of the corpus callosum: a rapid characterization of fiber bundle anatomy. *Front. Neuroanat*. 9:57. doi: 10.3389/fnana.2015.00057

Honey, C., Thivierge, J.-P., and OlSporns (2010). Can structure predict function in the human brain? *Neuroimage* 52, 766–776. doi: 10.1016/j.neuroimage.2010.01.071

Huang, S., Tobyne, S., Nummenmaa, A., Witzel, T., Wald, L., Mcnab, J., et al. (2016). Characterization of axonal disease in patients with multiple sclerosis using High-Gradient- Diffusion MR imaging. *Radiology* 280, 244–251. doi: 10.1148/radiol.2016151582

Huang, S. Y., Tian, Q., Fan, Q., Witzel, T., Wichtmann, B., McNab, J. A., et al. (2020). High-gradient diffusion MRI reveals distinct estimates of axon diameter index within different white matter tracts in the *in vivo* human brain. *Brain Struct. Funct*. 225, 1277–1291. doi: 10.1007/s00429-019-01961-2

Hursh, J. B. (1937). Conduction velocity and diameter of nerve fibers. *Am. J. Physiol. Legacy Content* 2, 382–390. doi: 10.1002/brb3.61

Innocenti, G. M., and Caminiti, R. (2017). Axon diameter relates to synaptic bouton size: structural properties define computationally different types of cortical connections in primates. *Brain Struct. Funct*. 222, 1169–1177. doi: 10.1007/s00429-016-1266-1

Innocenti, G. M., Caminiti, R., Rouiller, E. M., Knott, G., Dyrby, T. B., Descoteaux, M., et al. (2018). Diversity of cortico-descending projections: histological and diffusion MRI characterization in the monkey. *Cereb. Cortex* 29, 788–801. doi: 10.1093/cercor/bhx363

Jbabdi, S., and Johansen-Berg, H. (2011). Tractography: where do we go from here? *Brain Connect*. 1, 169–183. doi: 10.1089/brain.2011.0033

Jbabdi, S., Sotiropoulos, S. N., Haber, S. N., Van Essen, D. C., and Behrens, T. E. (2015). Measuring macroscopic brain connections in vivo. *Nat. Neurosci*. 18, 1546–1555. doi: 10.1038/nn.4134

Jelescu, I. O., Veraart, J., Fieremans, E., and Novikov, D. S. (2016). Degeneracy in model parameter estimation for multi-compartmental diffusion in neuronal tissue. *NMR Biomed*. 29, 33–47. doi: 10.1002/nbm.3450

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/s1361-8415(01)00036-6

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156.

Jeurissen, B., Descoteaux, M., Mori, S., and Leemans, A. (2017). Diffusion MRI fiber tractography of the brain. *NMR Biomed.* 42:e3785–22. doi: 10.1002/nbm.3785

Jeurissen, B., Leemans, A., Tournier, J.-D., Jones, D. K., and Sijbers, J. (2013). Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging. *Hum. Brain Mapp*. 34, 2747–2766. doi: 10.1002/hbm.22099

Jones, D. K., Alexander, D. C., Bowtell, R., Cercignani, M., Dell'Acqua, F., McHugh, D. J., et al. (2018). Microstructural imaging of the human brain with a 'super-scanner': 10 key advantages of ultra-strong gradients for diffusion MRI. *Neuroimage* 182, 8–38. doi: 10.1016/j.neuroimage.2018.05.047

Jones, D. K., and Cercignani, M. (2010). Twenty-five pitfalls in the analysis of diffusion MRI data. *NMR Biomed*. 23, 803–820. doi: 10.1002/nbm.1543

Jones, D. K., Knösche, T. R., and Turner, R. (2013). White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *Neuroimage* 73, 239–254. doi: 10.1016/j.neuroimage.2012.06.081

Jovicich, J., Czanner, S., Greve, D., Haley, E., Van Der Kouwe, A., Gollub, R., et al. (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *Neuroimage* 30, 436–443. doi: 10.1016/j.neuroimage.2005.09.046

Judson, M. C., Burette, A. C., Thaxton, C. L., Pribisko, A. L., Shen, M. D., Rumple, A. M., et al. (2017). Decreased axon caliber underlies loss of fiber tract integrity, disproportional reductions in white matter volume, and microcephaly in angelman syndrome model mice. *J. Neurosci.* 37, 7347–7361. doi: 10.1523/JNEUROSCI.0037-17.2017

Kakkar, L. S., Bennett, O. F., Siow, B., Richardson, S., Ianu, A., Quick, T., et al. (2018). Low frequency oscillating gradient spin-echo sequences improve sensitivity to axon diameter: an experimental study in viable nerve tissue. *Neuroimage* 182, 314–328. doi: 10.1016/j.neuroimage.2017.07.060

Lampinen, B., Szczepankiewicz, F., Mårtensson, J., van Westen, D., Sundgren, P. C., and Nilsson, M. (2017). Neurite density imaging versus imaging of microscopic anisotropy in diffusion MRI: a model comparison using spherical tensor encoding. *Neuroimage* 147, 517–531. doi: 10.1016/j.neuroimage.2016.11.053

Le Bihan, D., and Breton, E. (1985). Imagerie de diffusion *in vivo* par résonance magnétique nucléaire. *Comptes rendus de l'Acad. Sci.* 301, 1109–1112.

Le Bihan, D., and Iima, M. (2015). Diffusion magnetic resonance imaging: What water tells us about biological tissues. *PLoS Biol.* 13:e1002203. doi: 10.1371/journal.pbio.1002203

Lee, H. H., Jespersen, S. N., Fieremans, E., and Novikov, D. S. (2020a). The impact of realistic axonal shape on axon diameter estimation using diffusion MRI. *Neuroimage* 223:117228. doi: 10.1016/j.neuroimage.2020.117228

Lee, H. H., Papaioannou, A., Kim, S. L., Novikov, D. S., and Fieremans, E. (2020b). A time-dependent diffusion MRI signature of axon caliber variations and beading. *Commun. Biol.* 3:354. doi: 10.1038/s42003-020-1050-x

Lee, H. H., Yaros, K., Veraart, J., Pathan, J. L., Liang, F. X., Kim, S. G., et al. (2019). Along-axon diameter variation and axonal orientation dispersion revealed with 3D electron microscopy: implications for quantifying brain white matter microstructure with histology and diffusion MRI. *Brain Struct. Funct.* 224, 1469–1488. doi: 10.1007/s00429-019-01844-6

Morecraft, R. J., Binneboese, A., Stilwell-Morecraft, K. S., and Ge, J. (2017). Localization of orofacial representation in the corona radiata, internal capsule and cerebral peduncle in Macaca mulatta. *J. Comparat. Neurol.* 525, 3429–3457. doi: 10.1002/cne.24275

Moseley, M., Cohen, Y., Kucharczyk, J., Mintorovitch, J., Asgari, H., Wendland, M., et al. (1990). Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system. *Radiology* 176, 439–445. doi: 10.1148/radiology.176.2.2367658

Nakamura, H., and Kanaseki, T. (1989). Topography of the corpus callosum in the cat. *Brain Res.* 485, 171–175. doi: 10.1016/0006-8993(89)90679-3

Nilsson, M., Lasič, S., Drobnjak, I., Topgaard, D., and Westin, C. F. (2017). Resolution limit of cylinder diameter estimation by diffusion MRI: the impact of gradient waveform and orientation dispersion. *NMR Biomed.* 30:e3711. doi: 10.1002/nbm.3711

Novikov, D. S., Kiselev, V. G., and Jespersen, S. N. (2018). On modeling. *Magn. Res. Med.* 79, 3172–3193. doi: 10.1002/mrm.27101

Panagiotaki, E., Schneider, T., Siow, B., Hall, M. G., Lythgoe, M. F., and Alexander, D. C. (2012). Compartment models of the diffusion MR signal in brain white matter: A taxonomy and comparison. *Neuroimage* 59, 2241–2254. doi: 10.1016/j.neuroimage.2011.09.081

Paquette, M., Eichner, C., Knösche, T. R., and Anwander, A. (2020). Axon diameter measurements using diffusion mri are infeasible. *bioRxiv*. doi: 10.1101/2020.10.01.320507

Perge, J. A., Niven, J. E., Sterling, P., Mugnaini, E., and Balasubramanian, V. (2012). Why do axons differ in caliber? *J. Neurosci.* 32, 626–638. doi: 10.1523/JNEUROSCI.4254-11.2012

Pestilli, F., Yeatman, J. D., Rokem, A., Kay, K. N., and Wandell, B. A. (2014). Evaluation and statistical inference for human connectomes. *Nat. Methods* 11, 1058–1063. doi: 10.1038/nmeth.3098

Peters, A., Palay, S., and Webster, H. (1991). *The Fine Structure of the Nervous System: Neurons and Their Supporting Cells*. Oxford University Press.

Rafael-Patino, J., Romascano, D., Ramirez-Manzanares, A., Canales-Rodríguez, E. J., Girard, G., and Thiran, J. P. (2020). Robust monte-carlo simulations in diffusion-MRI: effect of the substrate complexity and parameter choice on the reproducibility of results. *Front. Neuroinform.* 14:8. doi: 10.3389/fninf.2020.00008

Rensonnet, G., Scherrer, B., Warfield, S. K., Macq, B., and Taquet, M. (2018). Assessing the validity of the approximation of diffusion-weighted-MRI signals from crossing fascicles by sums of signals from single fascicles. *Magn. Reson. Med.* 79, 2332–2345. doi: 10.1002/mrm.26832

Ritchie, J. (1982). On the Relation between Fibre Diameter and Conduction Velocity in Myelinated Nerve Fibres. *Proc. R. Soc. B Biol. Sci.* 217, 29–35.

Rudrapatna, U., Parker, G. D., Roberts, J., and Jones, D. K. (2021). A comparative study of gradient nonlinearity correction strategies for processing diffusion data obtained with ultra-strong gradient MRI scanners. *Magn. Res. Med.* 85, 1104–1113. doi: 10.1002/mrm.28464

Sherbondy, A. J., Dougherty, R. F., Ananthanarayanan, R., Modha, D. S., and Wandell, B. A. (2009). Think global, act local; projectome estimation with BlueMatter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Sherbondy, A. J., Rowe, M. C., and Alexander, D. C. (2010). "MicroTrack: an algorithm for concurrent projectome and microstructure estimation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Beijing).

Siow, B. M. C. W., Drobnjak, I., Ianus, A., Christie, I. N., Lythgoe, M. F., and Alexander, D. C. (2013). Axon radius estimation with oscillating gradient spin echo (OGSE) diffusion MRI. *Diffus. Fundam.* 18, 1–6.

Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2013). SIFT: spherical-deconvolution informed filtering of tractograms. *Neuroimage* 67, 298–312. doi: 10.1016/j.neuroimage.2012.11.049

Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2015). SIFT2: enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *Neuroimage* 119, 338–351. doi: 10.1016/j.neuroimage.2015.06.092

Szafer, A., Zhong, J., and Gore, J. C. (1995). Theoretical model for water diffusion in tissues. *Magn. Reson. Med.* 33, 697–712.

Tomasi, S., Caminiti, R., and Innocenti, G. M. (2012). Areal differences in diameter and length of corticofugal projections. *Cereb. Cortex* 22, 1463–1472. doi: 10.1093/cercor/bhs011

Tournier, J.-D., Calamante, F., and Connelly, A. (2007). Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage*, 35, 1459–1472. doi: 10.1016/j.neuroimage.2007.02.016

Tournier, J. D., Calamante, F., and Connelly, A. (2012). MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22, 53–66. doi: 10.1002/ima.22005

Van Gelderen, P., Despres, D., Vanzijl, P., and Moonen, C. (1994). Evaluation of restricted diffusion in cylinders. Phosphocreatine in rabbit leg muscle. *J. Magn. Reson. Ser. B* 103, 255–260.

Veraart, J., Fieremans, E., and Novikov, D. S. (2019). On the scaling behavior of water diffusion in human brain white matter. *Neuroimage* 185, 379–387. doi: 10.1016/j.neuroimage.2018.09.075

Veraart, J., Fieremans, E., Rudrapatna, U., Jones, D., and Novikov, D. (2018). "Breaking the power law scaling of the dMRI signal on the Connectom scanner reveals its sensitivity to axon diameters," in *Proceedings of the 27th Scientific Meeting of ISMRM* (Paris).

Veraart, J., Nunes, D., Rudrapatna, U., Fieremans, E., Jones, D. K., Novikov, D. S., et al. (2020). Noninvasive quantification of axon radii using diffusion MRI. *eLife* 9:e49855. doi: 10.7554/eLife.49855

Xu, J., Li, H., Harkins, K. D., Jiang, X., Xie, J., Kang, H., et al. (2014). Mapping mean axon diameter and axonal volume fraction by MRI using temporal diffusion spectroscopy. *Neuroimage* 103, 10–19. doi: 10.1016/j.neuroimage.2014.09.006

Xu, J., Li, H., Li, K., Harkins, K. D., Jiang, X., Xie, J., et al. (2016). Fast and simplified mapping of mean axon diameter using temporal diffusion spectroscopy. *NMR Biomed.* 29, 400–410. doi: 10.1002/nbm.3484

Zhang, H., Dyrby, T. B., and Alexander, D. C. (2011a). "Axon diameter mapping in crossing fibers with diffusion MRI," *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* (Toronto, ON).

Zhang, H., Hubbard, P., Parker, G., and Alexander, D. (2011b). Axon diameter mapping in the presence of orientation dispersion with diffusion MRI. *Neuroimage* 56, 1301–1315. doi: 10.1016/j.neuroimage.2011.01.084

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: Practical *in vivo* neurite orientation dispersion and density imaging of the human brain. *Neuroimage* 61, 1000–1016. doi: 10.1016/j.neuroimage.2012.03.072

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Zikopoulos, B., and Barbas, H. (2013). Altered neural connectivity in excitatory and inhibitory cortical circuits in autism. *Front. Hum. Neurosci.* 7:609. doi: 10.3389/fnhum.2013.00609

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 16 June 2021
doi: 10.3389/fnins.2021.661704

Multi-Channel 4D Parametrized Atlas of Macro- and Microstructural Neonatal Brain Development

Alena Uus[1], Irina Grigorescu[1], Maximilian Pietsch[2], Dafnis Batalle[2,3], Daan Christiaens[2,4], Emer Hughes[2], Jana Hutter[2], Lucilio Cordero Grande[2,5], Anthony N. Price[2], Jacques-Donald Tournier[2], Mary A. Rutherford[2], Serena J. Counsell[2], Joseph V. Hajnal[1,2], A. David Edwards[2] and Maria Deprez[1]*

[1] Department of Biomedical Engineering, School Biomedical Engineering and Imaging Sciences, King's College London, St. Thomas Hospital, London, United Kingdom, [2] Centre for the Developing Brain, School Biomedical Engineering and Imaging Sciences, King's College London, St. Thomas Hospital, London, United Kingdom, [3] Department of Forensic and Neurodevelopmental Science, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [4] Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium, [5] Biomedical Image Technologies, ETSI Telecomunicacion, Universidad Politécnica de Madrid, CIBER-BBN, Madrid, Spain

Structural (also known as anatomical) and diffusion MRI provide complimentary anatomical and microstructural characterization of early brain maturation. However, the existing models of the developing brain in time include only either structural or diffusion MRI channels. Furthermore, there is a lack of tools for combined analysis of structural and diffusion MRI in the same reference space. In this work, we propose a methodology to generate a multi-channel (MC) continuous spatio-temporal parametrized atlas of the brain development that combines multiple MRI-derived parameters in the same anatomical space during 37–44 weeks of postmenstrual age range. We co-align structural and diffusion MRI of 170 normal term subjects from the developing Human Connectomme Project using MC registration driven by both T2-weighted and orientation distribution functions channels and fit the Gompertz model to the signals and spatial transformations in time. The resulting atlas consists of 14 spatio-temporal microstructural indices and two parcellation maps delineating white matter tracts and neonatal transient structures. In order to demonstrate applicability of the atlas for quantitative region-specific studies, a comparison analysis of 140 term and 40 preterm subjects scanned at the term-equivalent age is performed using different MRI-derived microstructural indices in the atlas reference space for multiple white matter regions, including the transient compartments. The atlas and software will be available after publication of the article[1].

Keywords: multi-modal MRI, neonatal brain, spatio-temporal atlas, atlas-based analysis, multi-channel registration, white matter maturation, white matter parcellation

OPEN ACCESS

Edited by:
Tim B. Dyrby, Technical University of Denmark, Denmark

Reviewed by:
Frederik Lange, University of Oxford, United Kingdom
Viljami Sairanen, University of Verona, Italy

****Correspondence:***
Alena Uus
alena.uus@kcl.ac.uk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *31 January 2021*
Accepted: *20 May 2021*
Published: *16 June 2021*

Citation:
Uus A, Grigorescu I, Pietsch M, Batalle D, Christiaens D, Hughes E, Hutter J, Cordero Grande L, Price AN, Tournier J-D, Rutherford MA, Counsell SJ, Hajnal JV, Edwards AD and Deprez M (2021) Multi-Channel 4D Parametrized Atlas of Macro- and Microstructural Neonatal Brain Development. Front. Neurosci. 15:661704. doi: 10.3389/fnins.2021.661704

1. INTRODUCTION

In addition to being a routine diagnostic tool in neonatal brain imaging (Rutherford et al., 2010), MRI has been widely used for quantification and interpretation of neonatal brain development in term- and preterm-born infants. Premature birth before 37 weeks postmenstrual age (PMA) is associated with an increased risk of atypical brain maturation leading to neurocognitive and neurobehavioural disorders. Multiple studies demonstrated correlation of MRI metrics with

[1] 4D MC neonatal brain atlas: https://gin.g-node.org/alenaullauus/4d_multi-channel_neonatal_brain_mri_atlas

prematurity, clinical and environmental factors and neurodevelopmental outcomes (Ball et al., 2017; Barnett et al., 2018; Dimitrova et al., 2020). In this context, models of normal brain development such as spatio-temporal atlases (Schuh et al., 2018) can also potentially facilitate detection of altered maturation patterns. The advanced acquisition and reconstruction protocols (Cordero-Grande et al., 2018) produce high-resolution structural T1-weighted (T1w) and T2-weighted (T2w) MRI volumes that allow segmentation of fine brain anatomical structures (Makropoulos et al., 2014). But these MRI modalities have low contrast for white matter (WM) structures that also vary during the neonatal stage due to ongoing myelination. On the other hand, lower resolution diffusion MRI reflects the properties of tissue microstructural complexity in terms of diffusivity, anisotropy, neuronal density and fiber orientation (Pannek et al., 2012; Bastiani et al., 2019; Batalle et al., 2019; Feng et al., 2019; Pietsch et al., 2019; Zollei et al., 2019). Combined diffusion and structural MRI analysis has already shown a potential to increase interpretability of brain maturation patterns (Ball et al., 2017).

1.1. Structural MRI Metrics

The structural MRI-derived metrics most commonly used in neonatal brain studies include tissue- and structure-specific volumetry (Kuklisova-Murgasova et al., 2011; Makropoulos et al., 2016; Thompson et al., 2019) and surface measurements such as cortical thickness and curvature (Bozek et al., 2018; Fenchel et al., 2020) that can be extracted from automated segmentations (Makropoulos et al., 2014). Recently, automated segmentation of T2w images has also been applied for quantification of the volume of myelinated regions (Wang et al., 2019). Intensity changes in T1w and T2w images characterize white matter injury (O'Muircheartaigh et al., 2020) and diffuse excessive high signal intensity (DESHI) regions (Morel et al., 2021). Quantitative and semi-quantitative metrics applied to developing neonatal brains include the T1w/T2w signal ratio associated with myelin content (Bozek et al., 2018) and T2 relaxometry (Pannek et al., 2013; Kulikova et al., 2015; Wu et al., 2017; Knight et al., 2018).

1.2. Diffusion MRI Metrics

Brain microstructure can be probed using a variety of quantitative metrics derived from diffusion MRI. Even though diffusion tensor imaging (DTI) is limited by inconsistencies in fiber-crossing regions (Jeurissen et al., 2013), DTI-derived metrics, including the fractional anisotropy (FA) and the mean, radial and axial diffusitivity (MD, RD and AD) are still most widely used in neonatal brain studies (Barnett et al., 2018; Feng et al., 2019; Thompson et al., 2019; Dimitrova et al., 2020). Recently, higher order metrics, that alleviate some of the limitations of DTI in the fiber crossing regions, have also been applied to investigate neonatal brain development, including the mean kurtosis (MK) index derived from diffusion kurtosis imaging (DKI) (Bastiani et al., 2019) and intracellular volume fraction (ICVF), fiber orientation dispersion index (ODI) and volume fraction of the isotropic compartment (FISO) derived from Neurite Orientation Dispersion and Density Imaging (NODDI) model (Zhang et al., 2012). The NODDI-derived indices have been used to characterize development of both white and gray matter microstructural features (Kunz et al., 2014; Batalle et al., 2019; Fenchel et al., 2020; Kimpton et al., 2020). The microscopic fractional anisotropy (μFA) index (Kaden et al., 2016) designed to disentangle microscopic diffusion anisotropy from the orientation dispersion has not yet been applied to neonatal brains. Constrained spherical deconvolution (CSD) (Tournier et al., 2007; Jeurissen et al., 2014) allows extraction of orientation-resolved microstructural information as orientation distribution functions (ODF) from multi-shell high angular resolution diffusion imaging (HARDI) data. Based on fiber ODF, fixel-based analysis (Raffelt et al., 2017) provides the means for assessment of specific fiber populations in terms of fiber density (FD) and fiber-bundle cross-section (FC) (Pannek et al., 2018; Pecheva et al., 2019).

1.3. Atlases and Models of Neonatal Brain Development

Spatio-temporal normalization and construction of age-specific group-average templates have been routinely employed in processing pipelines in the recent large neonatal brain MRI studies to detect inter-group differences and anomalies in individual brains (Oishi et al., 2019). The majority of the reported spatio-temporal population-averaged atlases of the neonatal brain include either structural (T2w and T1w) (Kuklisova-Murgasova et al., 2011; Serag et al., 2012; Schuh et al., 2014, 2018; Wright et al., 2014; Makropoulos et al., 2016; Schwartz et al., 2016; Wang et al., 2019; O'Muircheartaigh et al., 2020) or diffusion (Feng et al., 2019; Pietsch et al., 2019; Dimitrova et al., 2020) channels. In this context, the term channel means an image of a single MRI contrast that is a part of a group of images belonging to the same subject. To our knowledge, the only existing multi-channel population-averaged 3D T1w+T2w+DTI atlas (Oishi et al., 2011) was constructed from a set of normal term subjects from 38 to 41 weeks PMA. However, the averaged template was reported to have significantly lower sharpness than the original T2w and DTI images. Apart from Feng et al. (2019) and Pietsch et al. (2019) who used FA+MD or multi-component ODF channels for registration, these atlases were constructed based on registration driven by a single channel and the output transformations were propagated to the rest. The reported multi-channel (MC) registration methods for brain studies are based on either combination of FA+structural (Park et al., 2003; Forsberg et al., 2011; Geng et al., 2012; Roura et al., 2015) or DTI+structural channels (Avants et al., 2007; Gupta et al., 2015; Irfanoglu et al., 2016). However, DTI-extracted metrics are characterized by inconsistencies in fiber-crossing regions (Tournier et al., 2012). In general, one of the challenges of multi-channel registration is considered to be the alignment between the structural and diffusion MRI volumes. Following spatial normalization, the templates are generally created using either weighted or direct averaging of the signal in the reference space. As an alternative, (Zhang et al., 2016) proposed to perform averaging in the frequency domain and reported higher sharpness of the atlas features.

Due to rapid changes of structure, volume and cytoarchitecture during the fetal and neonatal period, the majority of the atlases have also been resolved in time in the form of weekly templates. Smooth transitions between the atlas time points have been provided through kernel regression (Kuklisova-Murgasova et al., 2011; Serag et al., 2012; Schuh et al., 2014, 2018), logistic regression (Wang et al., 2019) or Gaussian process regression (Marquand et al., 2016; Dimitrova et al., 2020; O'Muircheartaigh et al., 2020). Recently, a Gompertz function (GF) was successfully used to parametrize fetal and neonatal brain volumetry and surface measurements (Wright et al., 2014; Makropoulos et al., 2016; Schwartz et al., 2016), showing better approximation than the linear model (Makropoulos et al., 2016), even though the changes in averaged structural (O'Muircheartaigh et al., 2020) and DTI (Bastiani et al., 2019; Feng et al., 2019; Dimitrova et al., 2020) metrics in white and gray matter can be approximated by linear trends. However, so far, there has been no reported works combining structural and diffusion MRI into a spatio-temporal atlas of the normal term born neonatal brain development.

1.4. Region Specific Analysis

The majority of neonatal brain studies have employed region-specific quantitative analyses based on correlation between the MRI-derived metrics measured within specific regions and parameters such as gestational age (GA) at birth, clinical factors or neurodevelopmental outcomes. In structural-only MRI datasets, segmentation is normally performed by atlas-based methods (Makropoulos et al., 2014). In the WM atlas-based analysis, the parcellation maps for the single-subject or population-average WM DTI atlases (Oishi et al., 2011; Feng et al., 2019; Alexander et al., 2020) were created by 2D manual delineation based on DTI directionally-encoded color maps for single subject or population-averaged templates. Label propagation based on DTI channel-guided registration has been widely used in neonatal brain studies (Kersbergen et al., 2014; Rose et al., 2014; Wu et al., 2017; Claessens et al., 2019; Feng et al., 2019). The tract-based spatial statistics (TBSS) (Smith et al., 2006) approach uses skeletonized FA maps for definition of the regions (Krishnan et al., 2016; Barnett et al., 2018; Young et al., 2018; Thompson et al., 2019). As an alternative, tract-specific analysis employs tractography to identify and segment the major WM pathways (Kulikova et al., 2015; Akazawa et al., 2016; Pecheva et al., 2017; Bastiani et al., 2019; Zollei et al., 2019; Dubner et al., 2020; Kimpton et al., 2020). In this case, the seed regions for tractography are defined in the template space and the segmentation of WM tracts is achieved by thresholding of the resulting probabilistic tractography maps. In Akazawa et al. (2016), this approach was also used to create population-specific average probabilistic maps of the major WM tracts.

1.5. Contributions

In this work, we propose to merge multiple metrics extracted from both diffusion and structural MRI in a single multi-channel spatio-temporal atlas of normal neonatal brain development parametrized using Gompertz function.

The generated 4D multi-channel atlas covers 37 to 44 weeks PMA range and includes structural (T1w, T2w and T1w/T2w myelin contrast) and diffusion channels with ODF, DTI, DKI, μFA and NODDI derived metrics. Furthermore, the atlas includes two parcellation maps: (i) the major WM tract regions (Alexander et al., 2020) refined using probabilistic tractography in the template space and (ii) a map of the transient WM regions associated with high maturation rates during the neonatal period. To ensure accuracy of spatial alignment, we propose MC registration method (Uus et al., 2020) guided by spatially-weighted structural MRI, diffusion (ODF) MRI and cortical segmentation (Makropoulos et al., 2018) channels. Parametrization in time is performed by the Gompertz function widely used for fitting of growth data. We implemented the atlas construction and fitting functionalities based on the MRtrix3 software package (Tournier et al., 2019). To demonstrate the application of the proposed atlas we perform a multi-modality study to compare term and preterm brain development and identify regions where WM maturation has been altered by preterm birth.

2. MATERIALS AND METHODS

2.1. Cohort, Datasets and Preprocessing

The atlas was constructed using 170 multi-modal MRI datasets of term-born neonates (born and scanned between 37 and 44 weeks PMA) that included T1w, T2w and HARDI scans. An additional 40 datasets of preterm neonates (born between 23 and 32 weeks GA: 28.94∓2.54 and scanned between 37 and 44 weeks PMA) were used for comparison analysis. Inclusion criteria were high image quality for scans of all modalities, singleton pregnancies and no major brain abnormalities. All scans were acquired under the developing Human Connectome Project (dHCP)[2]. The datasets were qualitatively assessed and graded by a team of dCHP researchers in terms of the reconstruction and motion correction quality, SNR levels, presence of artifacts and the global coverage of the brain ROI. Only the datasets with the best image quality were selected for this particular study. The distribution of the GA at birth and PMA at scan is given in **Figure 1**.

The datasets were acquired without sedation on a 3T Philips Achieva scanner equipped with a dedicated 32-channel neonatal head coil and baby transportation system (Hughes et al., 2017). The multi-shell HARDI volumes were acquired with four phase-encode directions on four shells with b-values of 0(20), 400(64), 1000(88) and 2,600(128) s/mm^2, TE 90 ms, TR 3800 ms (Hutter et al., 2018; Tournier et al., 2020) with $1.5 \times 1.5 \times 3$ mm resolution and 1.5 mm slice overlap and reconstructed to 1.5 mm isotropic resolution using the spherical harmonics and radial decomposition (SHARD) pipeline (Christiaens et al., 2018, 2021) that includes slice-wise motion correction, distortion correction and exclusion of corrupted slices. Prior to reconstruction, the diffusion datasets were preprocessed using the dedicated dHCP pipeline including: Marchenko-Pastur-PCA-based denoising (Veraart et al., 2016)

[2]dHCP project: http://www.developingconnectome.org.

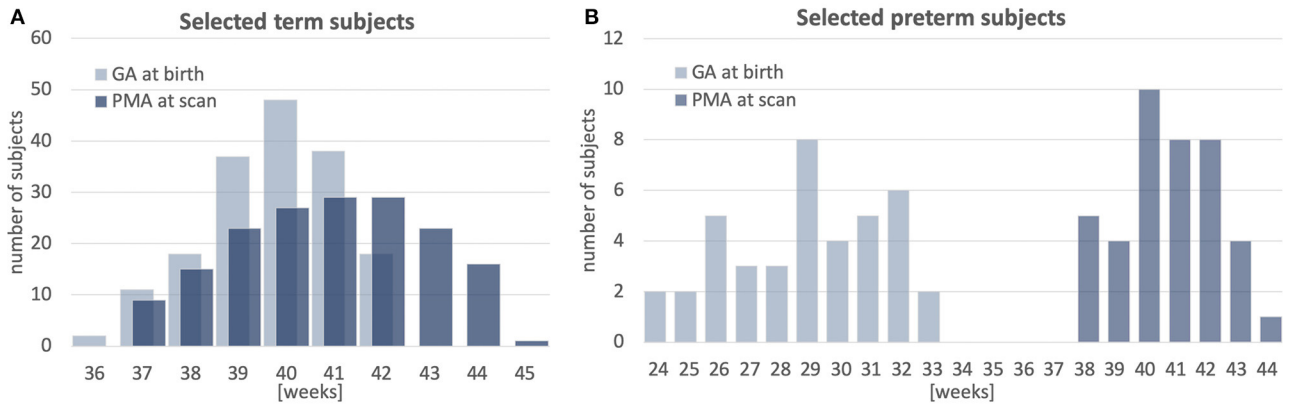


FIGURE 1 | Selected cohort of neonatal subjects from dHCP project: GA at birth and PMA at scan of 170 term subjects **(A)** and 40 preterm subjects **(B)**.

(MRtrix3[3]), Gibbs ringing removal (Kellner et al., 2016), susceptibility and eddy-current distortion correction and inter-volume motion correction with outlier replacement using topup (Andersson et al., 2003) (FSL[4]) and eddy (Andersson and Sotiropoulos, 2016) (FSL), bias field correction based on the $b = 0$ shell using N4 (Tustison et al., 2010) (ANTs[5]).

The structural T2w volumes were acquired using a TSE sequence with TR 12 *s*, TE 156 *ms*. The T1w volumes were acquired using an IR TSE sequence with TR 4.8 *s*, TE 8.7 *ms*. The isotropic T2w and T1w volumes with 0.5 *mm* resolution were reconstructed using a combination of motion correction (Cordero-Grande et al., 2018) and super-resolution reconstruction (Kuklisova-Murgasova et al., 2012). Intensities of individual T1w and T2w volumes were bias-corrected and normalized to the same intensity ranges as a part of the standard dHCP preprocessing pipeline based on DRAW-Em[6] (Makropoulos et al., 2014, 2018). In addition, the T2w images were normalized with respect to mean CSF signal intensity. The brain tissue and structure segmentations were generated by DRAW-Em pipeline (Makropoulos et al., 2014). For each dataset, the structural and diffusion volumes were coaligned based on affine registration of T2w and MD volumes using normalized cross-correlation (NCC) similarity metric implemented in MRTrix3. The diffusion-weighted imaging (DWI) volumes were globally normalized prior to the nonlinear multi-channel registration step (Tournier et al., 2019).

2.2. Extraction of MRI Metrics

The structural metrics include normalized T1w and T2w intensities and the T1w/T2w ratio reported to be associated with the myelin content (Glasser and Van Essen, 2011). Furthermore, we extracted Jacobians (J) of deformation fields from the MC registration output (section 2.4) to measure local volumetric changes.

The DTI metrics included MD, RD and FA extracted using MRtrix3 toolbox (Tournier et al., 2019). The DKI fitting and calculation of MK was performed similarly to Bastiani et al. (2019). The NODDI (Zhang et al., 2012) toolbox was used for fitting FISO, ICVF and ODI metrics. The estimation of micro FA maps was performed using SMT toolbox (Kaden et al., 2016). Only the two top HARDI shells were used for μFA and DKI fitting in order to minimize the impact of artifacts. In addition, we computed the mean DWI signal mDWI for the top 2, 600 s/mm^2 shell since it provides high contrast for WM structures. We extracted WM ODF from HARDI using MRtrix3 multi-shell multi-tissue constrained spherical deconvolution (Jeurissen et al., 2014). The track density imaging (TDI) maps were generated in the original space of dMRI volumes from the outputs of the standard MRtrix3 probabilistic tractography based on the 2nd order integration over fiber orientation distributions (iFOD2) (Tournier et al., 2010, 2019) with whole brain as the seed region and 700,000 streamlines for all datasets. This particular number of streamlines was selected arbitrarily.

2.3. Multi-Channel Registration of Blue Combined Structural and HARDI MRI Datasets

We propose a multi-channel non-linear registration technique to improve accuracy of spatial normalization of both structural and diffusion MRI images. The method is build on a multi-contrast ODF registration framework (Raffelt et al., 2011; Pietsch et al., 2017) implemented in MRtrix3 (Tournier et al., 2019) which employs SyN Demons (Avants et al., 2007) with an SSD metric and reorientation of ODF using apodized point spread functions (Raffelt et al., 2012). In order to decrease the sensitivity to acquisition or physiology related changes in signal intensities, we propose to replace the the standard SSD metric with a new robust local angular correlation (LAC) registration metric for ODF channels, which is an extension of angular correlation (Anderson, 2005) originally proposed for for quantitative assessment of ODF datasets. We further add structural and tissue parcellation channels with local NCC

[3]MRtrix3 toolbox: https://www.mrtrix.org.
[4]FSL toolbox: https://fsl.fmrib.ox.ac.uk.
[5]ANTs toolbox: http://stnava.github.io/ANTs.
[6]DRAW-Em toolbox: https://github.com/MIRTK/DrawEM.

(LNCC) similarity measure. The channels are combined through weighted fusion of the displacement field updates (Forsberg et al., 2011). Implementation of the LAC and LNCC metrics is based on the registration pipeline in MRtrix3 (Tournier et al., 2019) that includes reorientation of ODF (Raffelt et al., 2012).

In ODF diffusion model, diffusion signal is represented as a linear combination of real valued spherical harmonic (SH) orthonormal basis functions $Y_{lm}(\theta,\phi)$. For the task of image registration, two dMRI volumes can be expressed in terms of spatially varying spherical functions $A^{ODF}(\theta,\phi,x)$ and $B^{ODF}(\theta,\phi,x)$, where θ,ϕ are coordinates on the sphere and x is a spatial location:

$$A^{ODF}(\theta,\phi,x)=\sum_{l=0}^{\propto}\sum_{m=-l}^{l}a_{lm}(x)Y_{lm}(\theta,\phi)$$
$$B^{ODF}(\theta,\phi,x)=\sum_{l=0}^{\propto}\sum_{m=-l}^{l}b_{lm}(x)Y_{lm}(\theta,\phi) \quad (1)$$

We define **local angular correlation** r_a between A^{ODF} and B^{ODF} as:

$$r_a(x)=\frac{\langle A,B\rangle_x}{\langle A\rangle_x^{\frac{1}{2}}\langle B\rangle_x^{\frac{1}{2}}}=\frac{\sum_{x'\in N(x)}\sum_{l=2}^{L}\sum_{m=-l}^{l}a_{lm}(x')b_{lm}(x')}{(\sum_{x'\in N(x)}\sum_{l=2}^{L}\sum_{m=-l}^{l}a_{lm}^2(x'))^{\frac{1}{2}}(\sum_{x'\in N(x)}\sum_{l=2}^{L}\sum_{m=-l}^{l}b_{lm}^2(x'))^{\frac{1}{2}}}, \quad (2)$$

where A and B are 4D images of SH coefficients of order L with even $l=\{2,4,...,L\}$ harmonic degree terms, e.g., $A(x)=\{a_{lm}(x)\}_{l=2,...,L,m=-l,...,l}$ and $B(x)=\{b_{lm}(x)\}_{l=2,...,L,m=-l,...,l}$, $N(x)$ is the local neighborhood centered at x, and $<>_x$ denotes the inner product calculated over $N(x)$. $A(x)$ and $B(x)$ are also normalized with respect to local means (Avants et al., 2008). In this case, the $l=0$ term does not contribute to r_a values.

Since this is a correlation metric, the corresponding symmetric updates to the displacement fields Λ^A and Λ^B can be computed in a similar manner to LNCC demons (Avants et al., 2008):

$$\Lambda^A(x)=\frac{2\langle A,B\rangle_x}{\langle A\rangle_x\langle B\rangle_x}\left(B(x)-\frac{\langle A,B\rangle_x}{\langle A\rangle_x}A(x)\right)\nabla A(x) \quad (3)$$

$$\Lambda^B(x)=\frac{2\langle A,B\rangle_x}{\langle A\rangle_x\langle B\rangle_x}\left(A(x)-\frac{\langle A,B\rangle_x}{\langle B\rangle_x}B(x)\right)\nabla B(x)$$

Note that LAC operates in 4D (3D space plus SH dimension) while LNCC is calculated in 3D spatial neighborhood for each individual ODF channel separately (Raffelt et al., 2011).

In the proposed multi-channel registration pipeline, the fixed and moving inputs consist of a set of structural (e.g., T2w) and ODF channels $i=1,...,I$. At every iteration, the fixed A_i and moving B_i images are registered individually resulting in Λ_i^A and Λ_i^B updates to the displacement fields. The contributions from each of the channels to the global symmetric displacement field updates Λ_{MC}^A and Λ_{MC}^B are locally weighted by 3D gradient certainty maps based on the approach proposed in Forsberg et al. (2011).

First, at every iteration, the certainty gradient maps α_i^A and α_i^B are computed from the current version of warped channels A_i and B_i (including both structural and ODF volumes) and normalized as:

$$\alpha_i^A=\parallel\nabla A_i^T\nabla A_i\parallel,\quad \widehat{\alpha_i}^A=\frac{\alpha_i^A}{max(\alpha_i^A)} \quad (4)$$

Then, the global symmetric MC updates to the displacement fields Λ_{MC}^A and Λ_{MC}^B are computed by weighted averaging of the channel-specific update fields

$$\Lambda_{MC}^A=\frac{\sum_i\widehat{\alpha_i}^A\Lambda_i^A}{\sum_i\widehat{\alpha_i}^A},\quad \Lambda_{MC}^B=\frac{\sum_i\widehat{\alpha_i}^B\Lambda_i^B}{\sum_i\widehat{\alpha_i}^B} \quad (5)$$

This downweights the contributions of the regions in individual channels characterized by low contrast, ensuring that the output deformation fields are locally defined by the channels with the highest structural content. In comparison, the multi-channel SyN approach (Avants et al., 2007) or the existing alternative DTI-based MC registration methods (Geng et al., 2012; Gupta et al., 2015) employ simple averaging of the individual channel updates. **Figure 2** shows an example of certainty maps of T2w, ODF and cortex mask channels computed for one of the dHCP subjects along with the average MC weights used for normalization.

2.4. Generation of 4D Multi-Channel Atlas

The 4D parametrized MC atlas of neonatal brain development was generated from 170 term neonatal datasets in three sequential steps: (A) initial registration of structural channels to a single structural template and creation of an average multi-channel template, (B) refined registration of structural and diffusion channels to the multi-channel template and creation of age-dependent average multi-channel templates, (C) fitting of the signal and deformation fields in time using the Gompertz function to generate the parametrized 4D multi-channel atlas. The proposed pipeline is summarized in **Figure 3**.

2.4.1. Generation of a 3D Multi-Channel Template

We chose the T2w 36 week template from the dHCP neonatal brain atlas[7] (Schuh et al., 2018) as the global 3D reference space ($Y^{(reforg)}$) due to the lower degree of gyrification that facilitates more accurate registration of the cortex. All datasets $\{X_i\}_{,i=1,...,N}$ were registered to this template using affine alignment with global NCC followed by non-linear registration guided by two structural

[7]dHCP weekly neonatal brain atlas: https://gin.g-node.org/BioMedIA/dhcp-volumetric-atlas-groupwise.

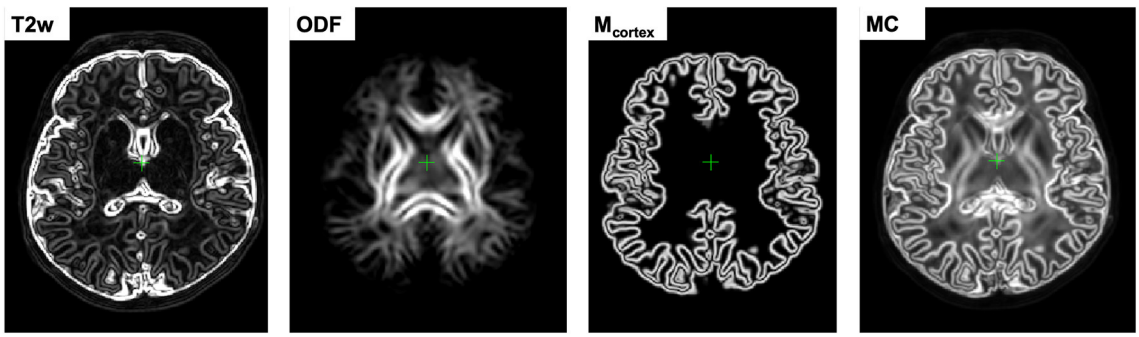


FIGURE 2 | An example of gradient-based certainty maps of T2w, ODF and cortex mask channels computed for one of the dHCP subjects along with the average MC gradient map used for normalization.

channels (T2w + cortex mask), similarly to O'Muircheartaigh et al. (2020):

$$W_i^{(1)} = \mathfrak{D}^{LNCC}(Y_c^{(reforg)}, X_{i,c}),\ _{c=\{T2;\ M_{cortex}\};i=1,...,N}, \quad (6)$$

where $\mathfrak{D}$ is the MC Demons registration operator, $W_i^{(1)}$ are the output deformation warps for each of the N datasets $X_{i,c}$ with $c = \{T2;\ M_{cortex}\}$ channels and $Y_c^{(ref)}$ is the reference volume. The MC registration included spatially weighted fusion of the channels (section 2.3, Uus et al., 2020). The output deformation warps $\{W_i^{(1)}\}_{,i=1,...,N}$ were propagated to the rest of the structural and dMRI channels. The preliminary set of 3D MC templates $\{Y_c^{(1)}\}_{,c=\{T2;\ M_{cortex};\ normODF\}}$ was generated by weighted averaging of all registered volumes of T2w, cortex mask and normalized (section 2.1) ODF channels (**Figure 3A**).

2.4.2. Generation of Age-Specific Multi-Channel Templates

At the second iteration (**Figure 3B**), we used registration with T2w + cortex mask + normalized ODF channels (section 2.3) to align all datasets to the multi-channel template (section 2.4.1):

$$W_i^{(2)} = \mathfrak{D}^{LNCC+LAC}(Y_c^{(1)}, X_{i,c}),\ _{c=\{T2;\ M_{cortex};\ normODF\};i=1,...,N} \quad (7)$$

Next, the datasets were divided into 15 subsets according to PMA, to sample the range from 37 to 44 weeks PMA into 0.5 week time-windows. Each of the subsets N^t contains 6-17 subjects depending on availability. The templates $Y_{c,t}^{(2)}$ for each of the metrics (c) described in section 2.2 were generated by robust weighted averaging of the metric maps $X_{i,c}$ transformed with $W_i^{(2)}$ in subsets $i \in N^t$:

$$Y_{c,t}^{(2)} = \sum_{i \in N^t} \omega_{i,c} \cdot \Theta(X_{i,c}, W_i^{(2)}) / \sum_{i \in N^t} \omega_{i,c},\ _{t=37,...,44}, \quad (8)$$

where Θ is the transformation operator, c is the list of all channels (see **Figure 3C**). The voxel-wise weights $\omega_{i,c}$ are binary maps with all values with $>$ 1.5 standard deviations from the mean set to zero. This minimizes the impact of outliers due to any abnormalities, artifacts or local misregistrations are excluded.

The templates $Y_{c,t}^{(2)}$ are biased toward 36 weeks reference space, therefore we calculate the transformations to remove this bias for each time-point. Since the registration is symmetric, it is acceptable to choose the inverse warps $(W_i^{(2)})^{-1}$ to create the transformation $W_{av,t}^{-1}$ from the age-specific average space to the global reference space:

$$W_{av,t}^{-1} = \sum_{i \in N^t} (W_i^{(2)})^{-1} / N^t,\ _{t=37,...,44} \quad (9)$$

Similarly, we create average inverse affine transformation $A_{av,t}^{-1}$ by selecting only the scaling and shearing components, followed by averaging and inverting.

2.4.3. Parametrized 4D Multi-Channel Atlas

In the final step, a continuous 4D spatio-temporal multi-channel model of the developing neonatal brain (**Figure 3C**) was constructed by fitting the Gompertz growth curves to the time-dependent average metric maps and transformations. We propose the following form of the Gompertz function since it allows interpretation of both growth rate (γ) and peak in time (τ):

$$G(t) = (\alpha - \delta) exp(-exp(-\gamma(\tau - t))) + \delta, \quad (10)$$

where t is the time point, α and δ control the upper and lower limits of $G(t)$, γ represents the growth rate and τ is the center point corresponding to the growth peak. The model was fitted to the time-dependent average metric maps $Y_{c,t}^{(2)}$ and transformations $W_{av,t}^{-1}, A_{av,t}^{-1}$ using least square minimization to produce continuous spatio-temporal maps in the reference space as well as average inverse transformations:

$$Y_c^{ref}(t) = G(\alpha^c, \delta^c, \gamma^c, \tau^c, t),\ _{t=[37;44]} \quad (11)$$

$$W^{-1}(t) = G(\alpha^W, \delta^W, \gamma^W, \tau^W, t),\ _{t=[37;44]} \quad (12)$$

$$A^{-1}(t) = G(\alpha^A, \delta^A, \gamma^A, \tau^A, t),\ _{t=[37;44]}, \quad (13)$$

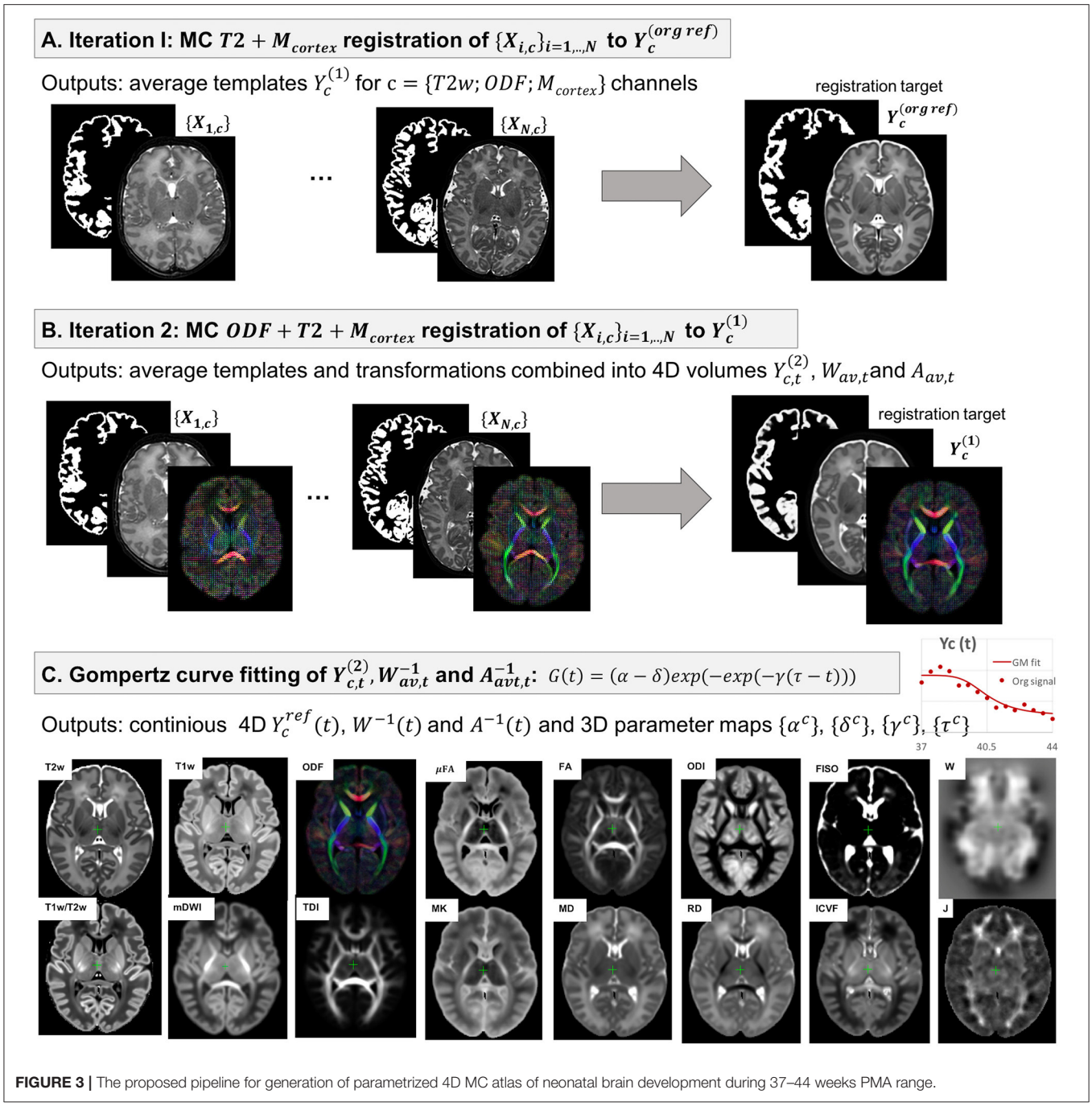


FIGURE 3 | The proposed pipeline for generation of parametrized 4D MC atlas of neonatal brain development during 37–44 weeks PMA range.

where α^c, δ^c, γ^c and τ^c are the Gompertz function parameters of metrics c = {T1w; T2w; T1w/T2w; mDWI; ODF: SH ODF, TDI; DTI: MD, RD, FA; DKI: MK; NODDI: ODI, FISO, ICVF; μFA; Jacobian} and t is continuous over 37–44 weeks PMA range. Unbiased spatio-temporal maps $Y_c(t)$ are obtained by applying nonlinear transformation $W^{-1}(t)$ followed by affine transformation $A^{-1}(t)$ to the biased spatio-temporal maps $Y_c^{ref}(t)$.

2.5. Parcellation of WM Regions

The dHCP structural atlas (Schuh et al., 2018) already provides parcellations of cortical and subcortical regions based on DRAW-EM pipeline (Makropoulos et al., 2014), therefore, this work specifically focuses on WM tracts and transient regions. At first, we propagated the parcellation map of the major WM tract regions from M-CRIB-WM atlas (a single subject template at 41 weeks PMA Alexander et al., 2020) by registration of one of the

T2w M-CRIB-WM atlas subjects to our T2w 44 week template $Y^{ref}_{T2w}(44)$.

Then we performed the MRTrix3 iFOD2 probabilistic tractography (Tournier et al., 2010) in $Y^{ref}_{ODF}(41)$ channel for each of the 54 WM regions (defined in Alexander et al., 2020) with propagated labels as seeds. We performed the tractography in the average template because of the lower noise levels due to averaging. This was followed by manual refinement of all labels using the 3D brush with thresholding editing tool in 3DSlicer (Fedorov et al., 2012) based on the thresholded TDI maps for individual tracts and inspection of the FA and T2 channels. The procedure was performed in three iterations with iFOD2 tractography being performed for the WM ROIs refined in the previous step. The labels were created in the atlas reference space resampled to 0.5 *mm* isotropic resolution to account for finer WM structures.

The transient WM regions were localized as regions with high rates of signal changes during 37–44 weeks PMA. The parcellation was generated semi-automatically from the γ^{av} map obtained by averaging the absolute of growth rate γ^{c} maps for T1w, T2w, RD and FISO channels. These channels were selected since they showed similar patterns in the region associated with the transient fetal compartments (Pittet et al., 2019). The γ^{av} map (with values varying within [0; 0.5]) was thresholded at 0.25 and manually refined.

2.6. Atlas-Based Region-Specific Analysis

In order to assess the feasibility of the proposed approach for atlas-based region-specific analysis studies, we performed a comparison of term and preterm cohorts. The analysis was based on both the WM and γ^{av} parcellation maps. At first, all subjects (selected 40 preterm and 140 term subjects scanned between 38 and 43 weeks PMA range) were registered to the PMA-matched atlas space (section 2.3) with T2w, ODF, cortex and ventricle mask channels. It was identified experimentally, that adding the ventricle mask channel improves registration results for preterm subjects since preterm brains commonly have enlarged ventricles. Therefore, it was used for all subjects in the term-preterm comparison study.

The comparison analysis between the cohorts was performed in the atlas space. The structural and dMRI metrics were computed for each of the ROIs using robust weighted averaging with only the values with the difference < 1.5 standard deviations from the mean included. The robust averaging helps to avoid errors due to image artifacts or local misregistration at the structure boundaries. The associations between the extracted metrics and the PMA at scan and the GA at birth were assessed using the standard ANOVA linear model analysis. The output p-values were corrected for multiple comparisons using the Bonferroni correction.

2.7. Implementation Details

The atlas was constructed with isotropic resolution 0.75 *mm*. The LAC metric for MC registration of ODF channels was implemented in MRtrix3 (Tournier et al., 2019). In addition, we implemented the LNCC Demons metric (Avants et al., 2008) in MRtrix3 for registration of the structural channels which, although described in Raffelt et al. (2011), was not available in the current implementation of MRtrix3. We chose the default MRtrix3 registration parameters[8] for multi-resolution ($\{0.5; 0.75; 1.0\}$), SH order ($l_{max} = \{0; 2; 2\}$), regularization of the gradient update field with Gaussian smoothing with 1 voxel standard deviation and regularization of the displacement field with Gaussian smoothing with 0.75 voxel standard deviation. For LNCC and LAC we chose the local neighborhood with 3 voxel radius (similarly to Raffelt et al., 2011). The proposed 4D GF fitting step (10) was implemented in MRtrix3. The ANOVA analysis for comparison between the term and preterm subjects was performed in RStudio (RStudio Team, 2020) using the standard `lm()` function.

3. RESULTS AND DISCUSSION

3.1. Multi-Channel Registration

In our previous work (Uus et al., 2020) we have demonstrated that the proposed MC registration improves overall alignment of cortical and WM regions when driven by both structural and ODF channels in longitudinal cases. Here we confirm these results in cross-sectional registration. Additionally, we demonstrate that including the cortex mask as an additional channel improves accuracy of cortical alignment, which is otherwise decreased in the presence of ODF channel. This approach was also used in Makropoulos et al. (2018) and O'Muircheartaigh et al. (2020) to improve single-channel T2w registration.

We investigated six scenarios of registration of individual dHCP subjects to the templates $Y^{ref}_{c}(t)$ based on different combinations of channels: (I) $T2w$, (II) $T2w + M_{cortex}$, (III) $T2w + M_{cortex} + FA$, (IV) $T2w + M_{cortex} + ODF(LAC)$, (V) $T2w + ODF(LAC)$ and (VI) $ODF(LAC)$. The performance was tested on 11 term datasets from 42.00 to 42.57 weeks PMA since at this age the subjects have significantly higher degree of gyrification than the average templates. To assess the alignment in both WM and cortical regions we evaluated similarity of aligned individual images with the age- and contrast-matched templates using mutual information (MI) for (A) T1w channel in the cortical region and (B) TDI channel in the dilated WM region (highlighted in yellow in **Figure 4**). The mutual information similarity metric and the T1w and TDI channels were selected for evaluation to minimize bias toward the channels and similarity metrics used in registration.

We observed that all ODF-guided scenarios led to highest quality alignment of TDI maps ($p < 0.001$) and adding additional channels did not decrease the similarity after alignment ($p > 0.05$). Including the FA channel improved TDI similarity compared to T2w and T2w+M ($p < 0.001$), but it was still significantly lower than for ODF guided alignments ($p < 0.001$) due to the contrast of poorly defined cortical features in FA. In the cortical region similarity of T1w contrast for the proposed $T2w + M_{cortex} + ODF$ MC registration was only slightly lower

[8]MRtrix3 mrregister function: https://mrtrix.readthedocs.io/en/latest/reference/commands/mrregister.html.

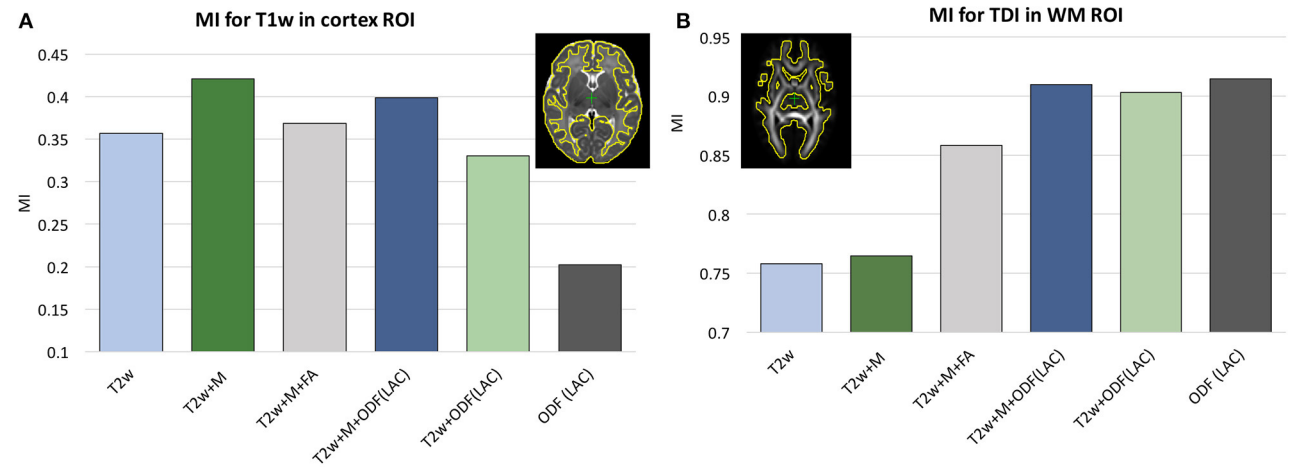


FIGURE 4 | Comparison of MC registration results for different combinations of channels. The performance was measured by mutual information (MI) between aligned images and the age- and contrast-mached templates for **(A)** the T1w images in the cortical region and **(B)** TDI maps in the WM region. The regions are highlighted in yellow contours. The results are statistically significant with $p < 0.001$ for all cases apart from: all *ODF*-guided scenarios for the WM ROI, $T2w$ vs. $T2w + M_{cortex}$ for WM ROI and $T2w$ vs. $T2w + M_{cortex} + FA$ for the cortex ROI.

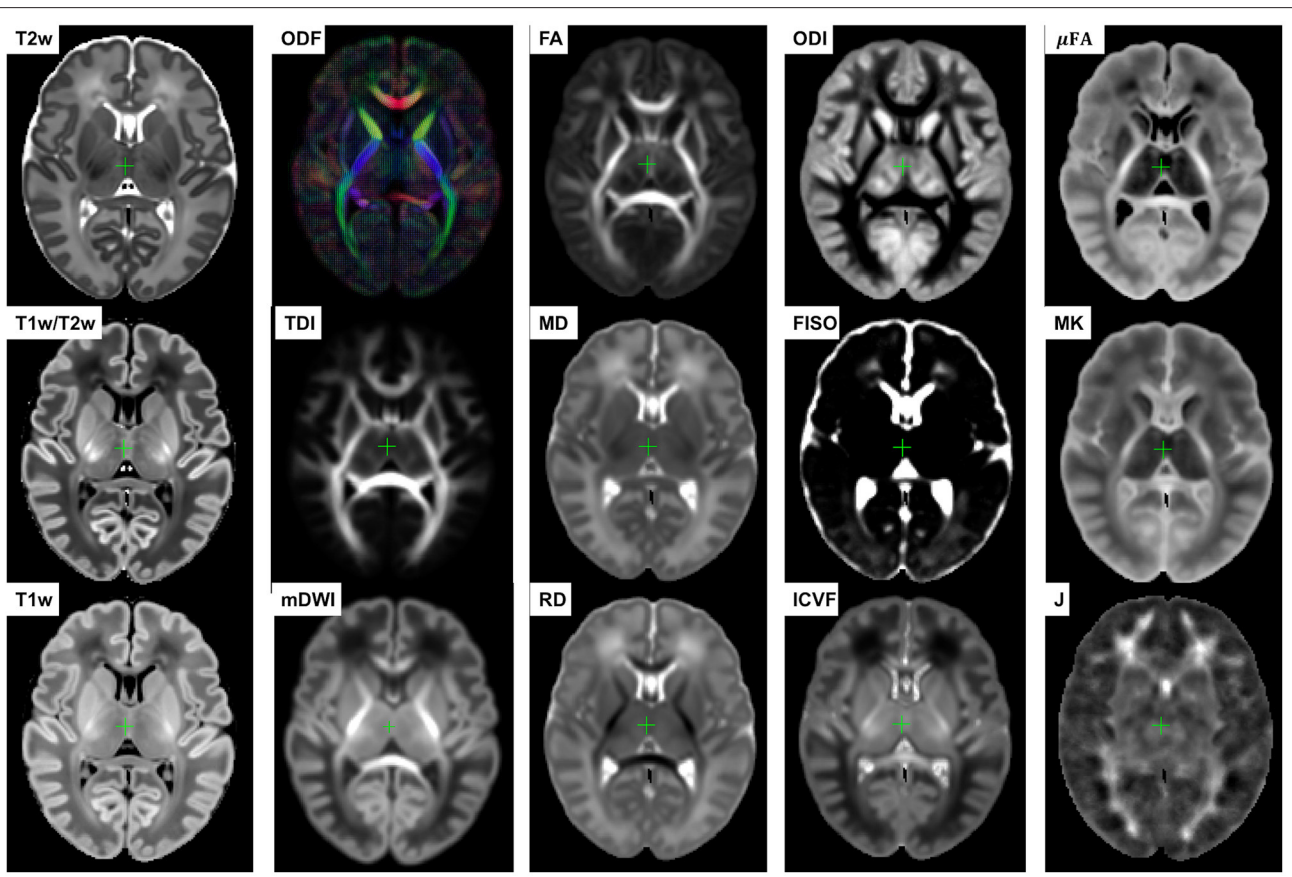


FIGURE 5 | Multi-channel 4D atlas in the reference space (corresponding to 36 weeks PMA). Structural channels: T1, T2, T1/T2 and Jacobian; ODF channels: SH ODF, mDWI, TDI; DTI channels: MD, RD, FA; DKI channel: MK; NODDI channels: ODI, FISO, ICVF; μFA.

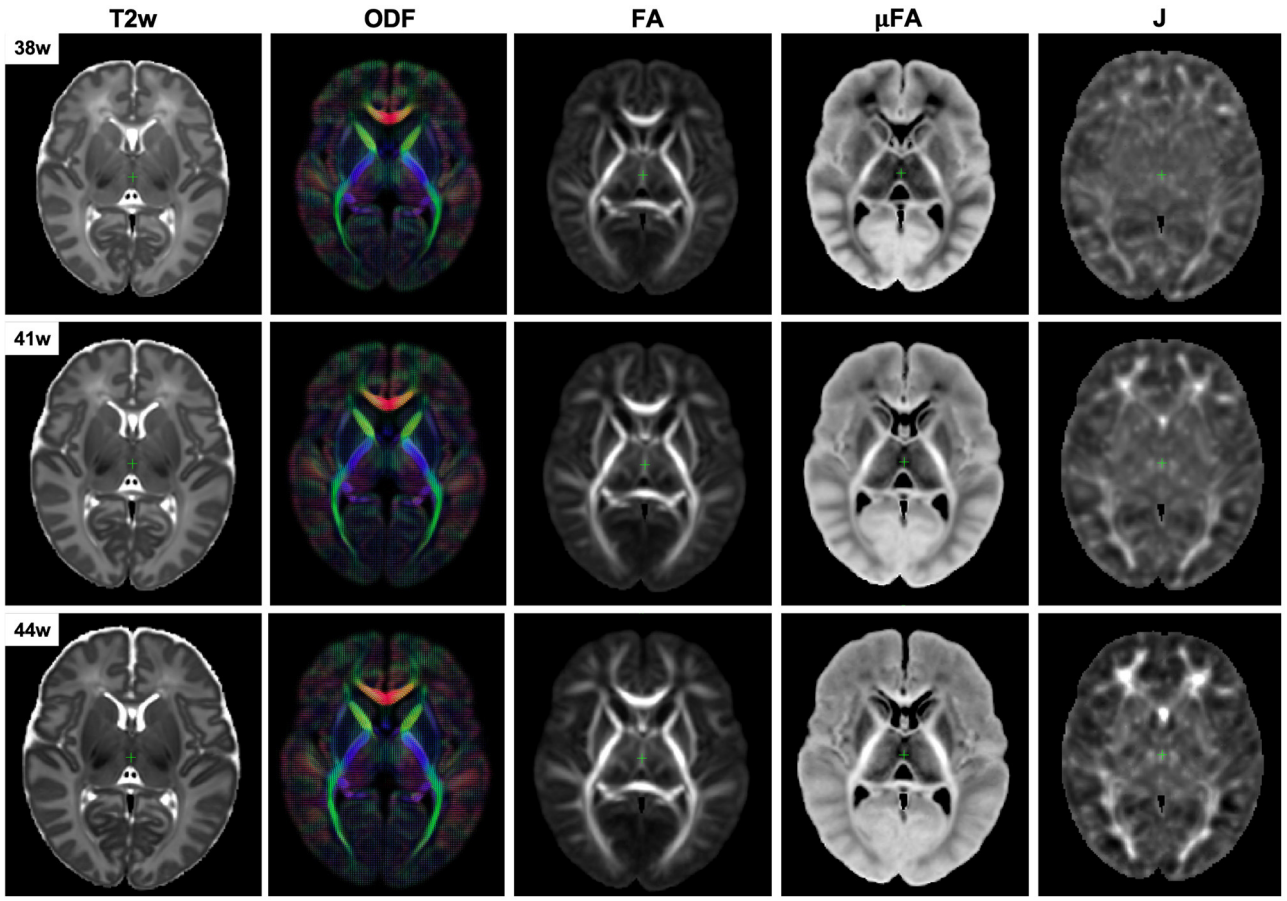


FIGURE 6 | Example unbiased 4D atlas channels at 38, 41, and 44 weeks PMA. The corresponding Jacobian maps (J) are shown in the reference space.

than the $T2w + M_{cortex}$, but it was significantly higher than all the other scenarios ($p < 0.001$). Addition of the M_{cortex} channel improved the cortical alignment in all cases thus resolving the limitation reported in our previous work (Uus et al., 2020).

3.2. 4D Multi-Channel Atlas of Normative Neonatal Brain Development

The resulting multi-channel 4D atlas $Y_c^{ref}(t)$ in the reference space (36 weeks PMA dHCP atlas Schuh et al., 2018) is shown in **Figure 5**. Unbiased atlases $Y_c(t)$ obtained after application of average inverse warps for 38, 41 and 44 weeks PMA time points are presented in **Figure 6**. There are distinct nonlinear changes due to cortical folding in the T2w templates and volumetric expansion/contraction due to growth the is visible in the Jacobian maps.

The created WM parcellations map with 54 ROIs created in the atlas reference space (section 2.5) for the region-specific analysis of the metric values is shown in **Figure 7B**. The label annotation information follows the original annotations defined in Alexander et al. (2020). The tractography-based manual refinement of the originally propagated 2D-slice-wise segmentations (**Figure 7A**) from the M-CRIB-WM atlas provided a more accurate 3D definition of the WM ROIs that are developed by 44 weeks PMA. Furthermore, it removed the structural inconsistencies in the original 2D slice-wise WM segmentations that were performed on DTI directionally-encoded color maps.

Figure 8A presents the parcellation map of the transient regions identified by high rates of signal changes during 37–44 weeks PMA segmented from the average γ^{av} map (**Figure 8B**). The parcellation map has 24 left/right regions with the majority being consistent with the transient fetal compartment regions described in the recently introduced extended MRI scoring systems of neonatal brain maturation (Pittet et al., 2019) including periventricular crossroads (Judaš et al., 2005), Von Monakow WM segments and subplate. We also identified fast developing regions within the cerebellum and subcortical gray matter.

In addition, we calculated voxel-wise R^2 scores to evaluate the Gompertz function fit. Our results confirmed that GF offers higher R^2 scores than linear regression with $p<0.001$ for the combined γ and WM parcellation map region. The primary

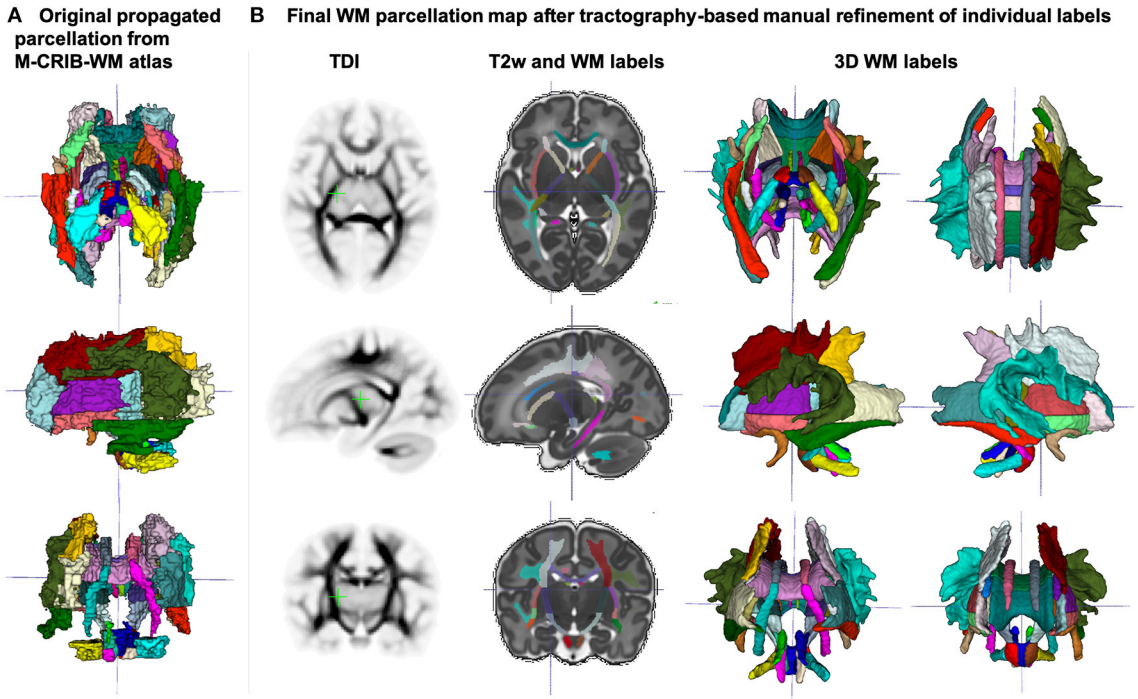


FIGURE 7 | (A) Original WM parcellation map propagated from the M-CRIB-WM atlas using T2w-guided registration. **(B)** Final WM parcellation map after tractography-based manual refinement in the atlas reference space. The 54 ROIs are based on the structures defined in the M-CRIB-WM atlas (Alexander et al., 2020). The corresponding TDI map highlights the WM pathway regions.

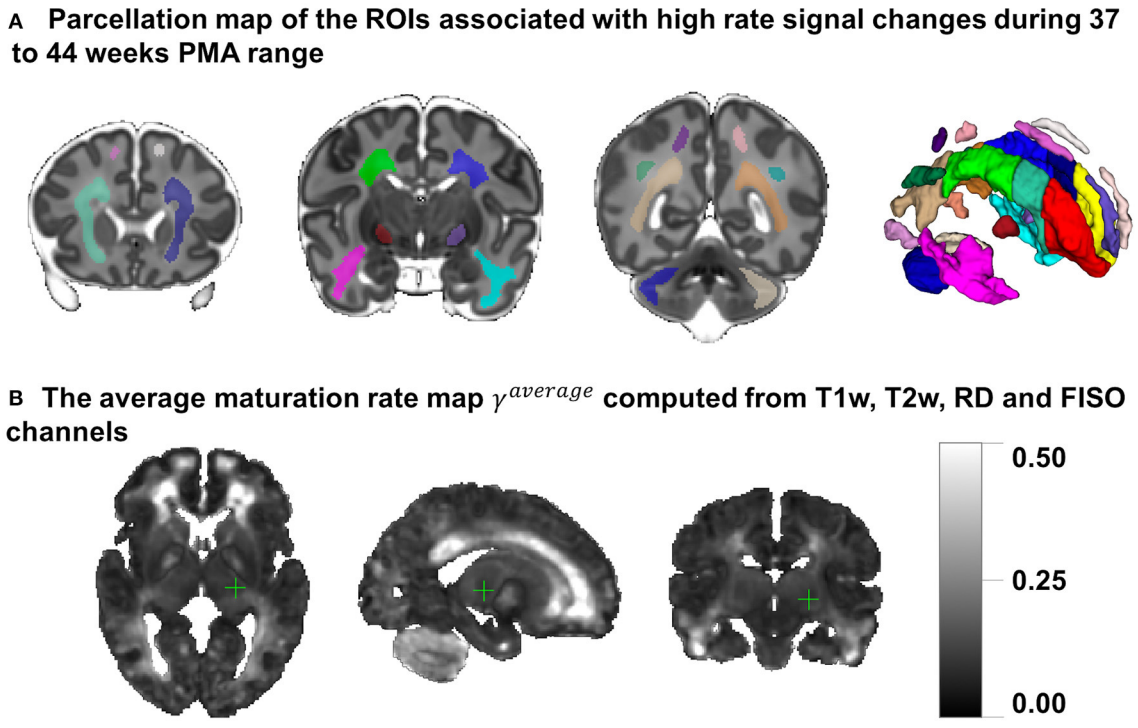


FIGURE 8 | (A) The parcellation map of 24 paired regions identified by high change rates during 37–44 week PMA. **(B)** The average maturation rate map γ^{av} computed from T1w, T2w, RD, and FISO channels.

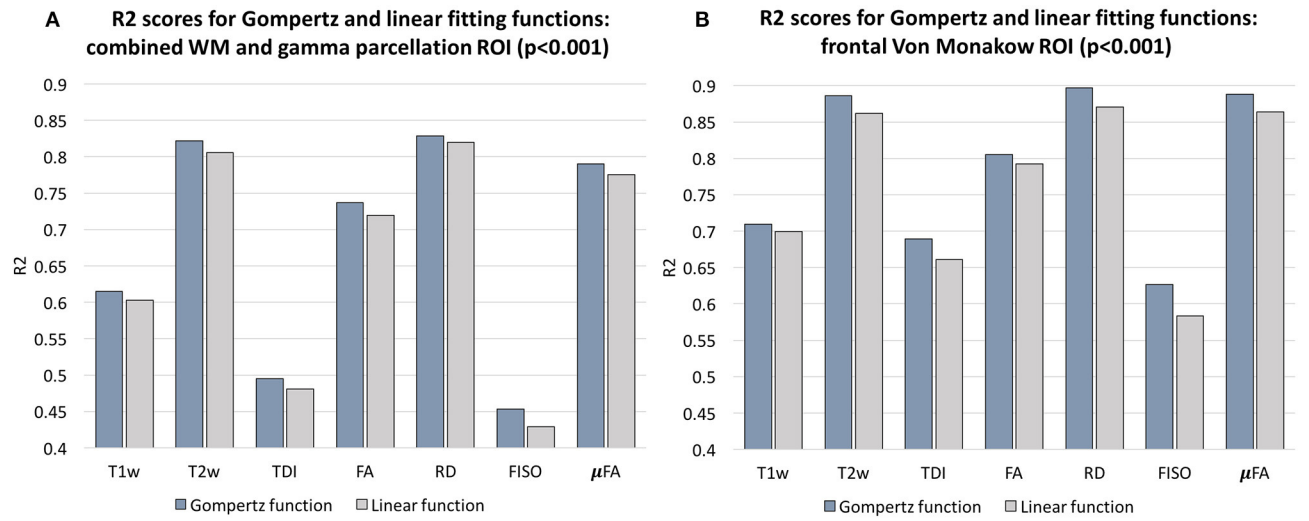


FIGURE 9 | Comparison of the Gompertz function and linear fitting results in terms of R^2 values evaluated within the combined WM and γ^{av} parcellation map regions **(A)** and frontal Von Monakow WM regions **(B)**. The results are statistically significant with $p < 0.001$.

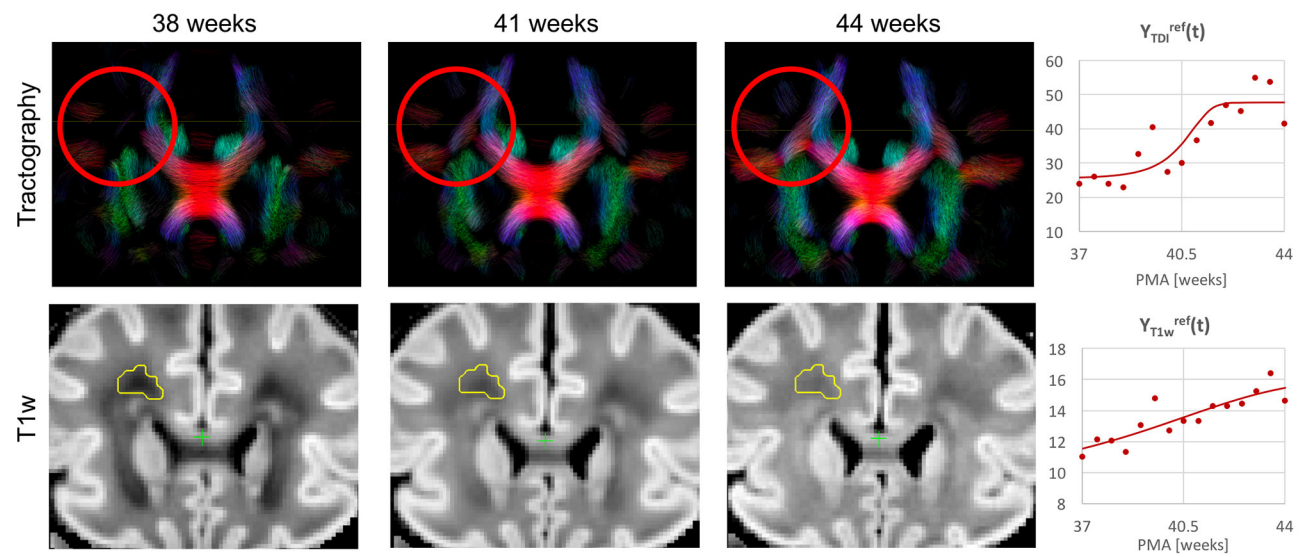


FIGURE 10 | Whole brain probabilistic tractography generated from the ODF channel $Y_{ODF}^{ref}(t)$ and the corresponding T1w channel $Y_{T1w}^{ref}(t)$ (in the reference space) in the frontal WM region at 38, 41, and 44 weeks PMA time points. The developing WM pathway (red circle) can be linked to the increasing T1w signal intensity (yellow region). The graphs show the signal in age-specific templates $Y_{c,t}^{(2)}$ and fitted Gompertz function $Y_c^{ref}(t)$ in the TDI and T1w channels averaged over the region highlighted in yellow.

regions where the GF fitting outperformed linear fitting were the γ^{av} parcellation map and the local WM regions such as the frontal Von Monakow WM regions (labels 1 and 4 in the γ^{av} parcellation map). **Figure 9** shows R^2 values for GF vs. linear fitting comparison for a subset of channels.

Examples of the non-linear patterns in signal changes also can be observed in the graphs in **Figures 10–13** showing average signal values in 3 × 3 × 3 voxel ROIs and the corresponding average GF fitting results. However, the relatively small improvement in R^2 suggests that a linear fit also offers a reasonable approximation during this short time-window and that it is acceptable to use the linear model based ANOVA analysis for interpretations of trends in early neonatal brain development.

3.3. Visual Analysis of Normal Neonatal Brain Development

Figure 10 shows the output of iFOD2 probabilistic tractography (Tournier et al., 2010) generated from the ODF channel and

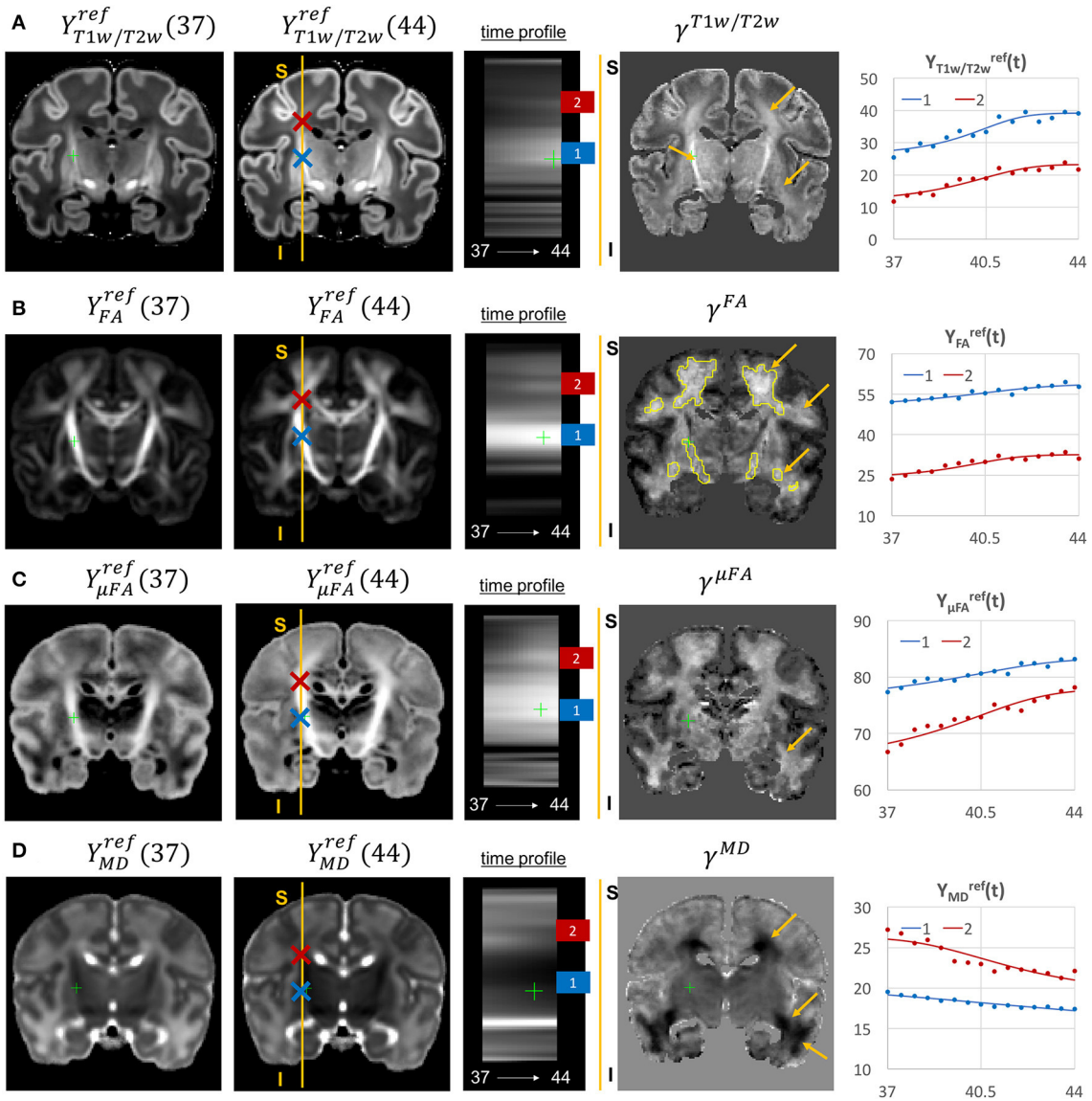


FIGURE 11 | Examples of the signal changes in time (in the reference space) in T1w/T2w **(A)**, FA **(B)**, μFA **(C)**, and MD **(D)** channels. First column: 37 week template. Second column: 44 week template. Third column: signal change in time. Fourth column: γ^c maps. Fifth column: Signal change in time in age-specific templates $Y^{(2)}_{c,t}$ and fitted Gompertz function $Y^{ref}_c(t)$ computed over 3 × 3 × 3 voxel regions in two locations: PLIC (blue) and superior corona radiata (red). The regions highlighted with yellow contours have > 0.2 weeks growth peak offset in τ^c.

the corresponding T1w channel (in the reference space) in the frontal WM region at 38, 41, and 44 weeks PMA time points. The increase in the T1w signal (known to be sensitive to proliferation of cells and myelin precursors and decreasing water content Girard et al., 2012) can be linked to the developing WM pathways seen in tractography (highlighted in red circle). The graphs show the corresponding increasing intensities in the age-specific average templates $Y^{(2)}_{c,t}$ and fitted signal values $Y^{ref}_c(t)$ of the TDI and T1w channels computed in the small frontal Von Monakow WM segment (Pittet et al., 2019) highlighted in yellow in the T1w channel.

The examples of signal intensity changes in time in different channels and the corresponding growth rate maps γ^c are presented in **Figures 11–13**. The regions highlighted in yellow have a growth peak offset in time $\geq$ 0.2 weeks from the 40.5 weeks central time point in τ^c and can be interpreted as indicators of earlier or later maturation with respect to the central time point of 40.5 weeks PMA. The graphs show average signal values in 15 discrete age-specific templates $Y^{(2)}_{c,t}$ and the corresponding fitted signal $Y^{ref}_c(t)$ calculated within small 3 × 3 × 3 voxel regions at specific locations, including the right posterior limb of internal capsule (PLIC), superior

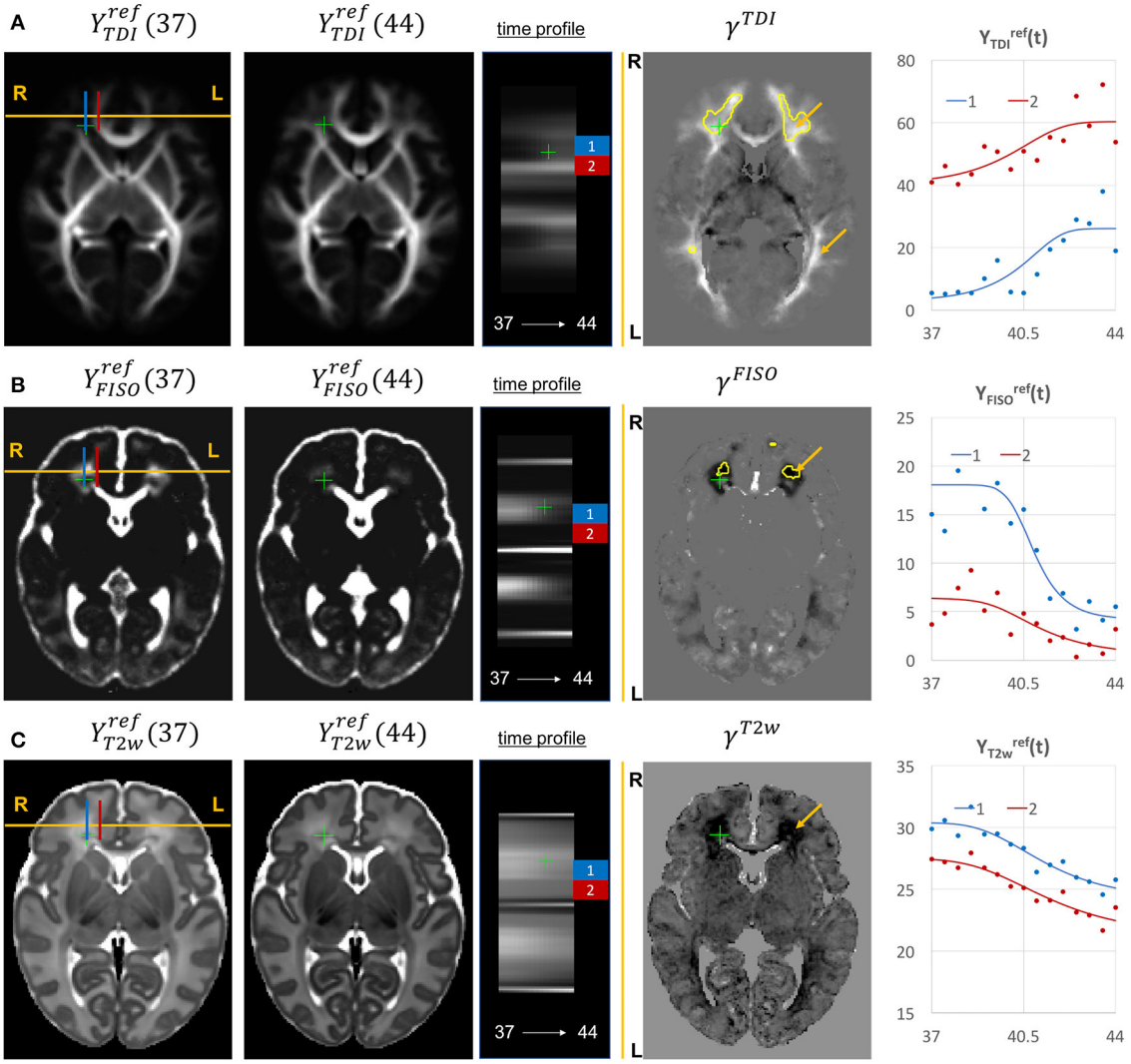


FIGURE 12 | Examples of the signal changes in time (in the reference space) in TDI **(A)**, FISO **(B)**, and T2w **(C)** channels. First column: 37 week template. Second column: 44 week template. Third column: signal change in time. Fourth column: γ^c maps. Fifth column: Signal change in time in age-specific templates $Y_{c,t}^{(2)}$ and fitted Gompertz function $Y_c^{ref}(t)$ computed over 3 × 3 × 3 voxel regions in two locations: prefrontal corpus callosum (red) and Von Monakow WM segment (blue). The regions highlighted with yellow contours have > 0.2 weeks growth peak offset in τ^c.

corona radiata, periventricular crossroads, corpus callosum, Von Monakow WM segment and cerebellum.

The WM tracts are characterized by different maturation times and rates (Iida et al., 1995). The T1w/T2w contrast (linked to myelination by Glasser and Van Essen, 2011) shows gradual signal increase from 37 to 44 weeks (**Figure 11A**). The $\gamma^{T1w/T2w}$ map and the average signal graphs $Y_{T1w/T2w}(t)$ confirm that the rate of T1w/T2w signal increase is the highest in the PLIC region (blue) and the corona radiata (red). The value of the $\tau^{T1w/T2w}$ parameter of the Gompertz function is approximately 40.5 weeks in both regions which is in agreement with the previously reported myelination milestones (Counsell et al., 2002; Wang et al., 2019). There is also a noticeable increase in the cortical T1w/T2w signal, also previously reported by Bozek et al. (2018), which may be due to the ongoing myelination or the increased cell density (Girard et al., 2012). Both FA and μFA signals (**Figures 11B,C**) gradually increase in all WM regions in agreement with the trends reported in Feng et al. (2019) and Dimitrova et al. (2020). The μFA map shows generally higher degree of changes than FA, potentially due to the increasing crossing fiber effect, while in γ^{FA}, the more prominent WM changes are observable primarily in the corona radiata, sagittal stratum and superior longitudinal fasciculus as well as the parietal crossroads and subplate (highlighted with arrows). The γ^{MD} map of the MD channel (**Figure 11D**) shows a large decrease in the superior corona radiata, sagittal stratum and the transient fetal compartments associated with WM maturation (Judaš et al., 2005; Pittet et al., 2019) including the periventricular crossroads

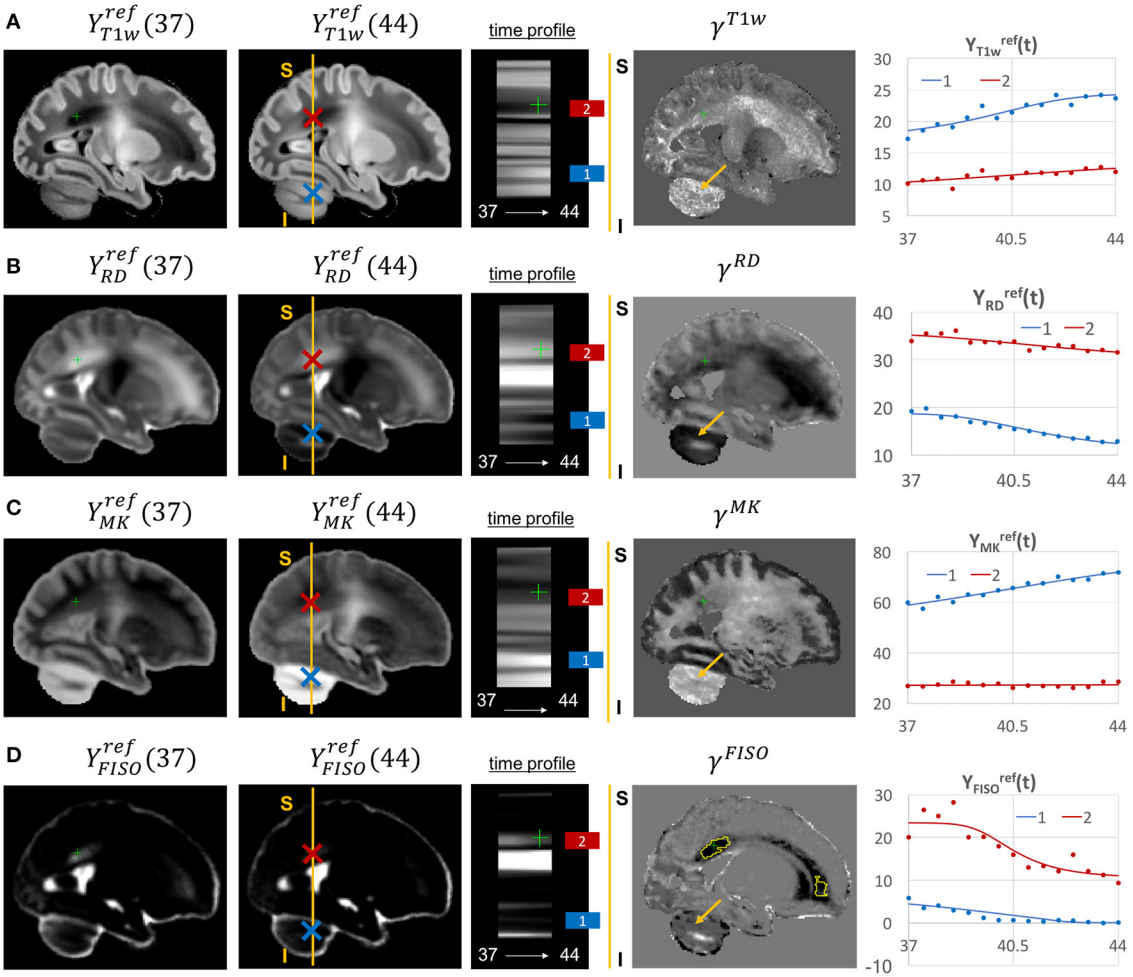


FIGURE 13 | Examples of the signal changes in time (in the reference space) in T1w **(A)**, RD **(B)**, MK **(C)**, and FISO **(D)** channels. First column: 37 week template. Second column: 44 week template. Third column: signal change in time. Fourth column: γ^c maps. Fifth column: Signal change in time in age-specific templates $Y^{(2)}_{c,t}$ and fitted Gompertz function $Y^{ref}_c(t)$ computed over 3 × 3 × 3 voxel regions in two locations: cerebellum (blue) and periventricular crossroads (red). The regions highlighted with yellow contours have > 0.2 weeks growth peak offset in τ^c.

and subplate regions (highlighted with arrows). The MD signal is slowly decreasing the PLIC region as can be seen in the corresponding graph (blue). All of the presented γ^c maps also show significant changes in the periventricular parietal crossroad regions (highlighted with arrows) with the significant decrease in MD and increasing in T1w/T2w.

Given the fixed number of streamlines used for probabilistic tractography, there is a notable redistribution of the TDI amplitude from the main to proximal WM tracts (**Figure 12A**). The corresponding growth rate γ^{TDI} map is positive in the frontal (anterior corona radiata) and thalamic radiation WM regions (highlighted with arrows) and negative in the internal capsule. The R-L time profile in the frontal region (Von Monakow WM segment, blue) shows the increased track density at 44 weeks. The average TDI signals $Y^{ref}_{TDI}(t)$ in this region (blue) and the corpus callosum (red) are also characterized by a significant degree of nonlinearity. the NODDI FISO component (**Figure 12B**) shows a prominent reduction in the same frontal region which is in agreement with the expected decrease of water content and progressing maturation of WM pathways (Girard et al., 2012). Similarly to TDI, the average FISO signals $Y^{ref}_{FISO}(t)$ in the investigated WM ROIs have nonlinear shape with the steep decrease occurring during the 39.5–43 weeks period. A similar decrease is observed in T2w signal (**Figure 12C**). The FISO channel in the sagittal view in **Figure 13D** also demonstrates similar patterns in the periventricular crossroads (red).

Most of the channels also show prominent changes in the cerebellum associated with the normal maturation process (**Figure 13**, blue). The T1w signal intensity $Y^{ref}_{T1w}(t)$ is gradually increasing due to WM development along with the increasing microstructural complexity reflected in the MK channel with the high γ^{MK} map values and the expected decreasing trends of the RD $Y^{ref}_{RD}(t)$ and FISO $Y_{FISO}(t)$ signals (potentially due to the decreasing amount of free water Girard et al., 2012).

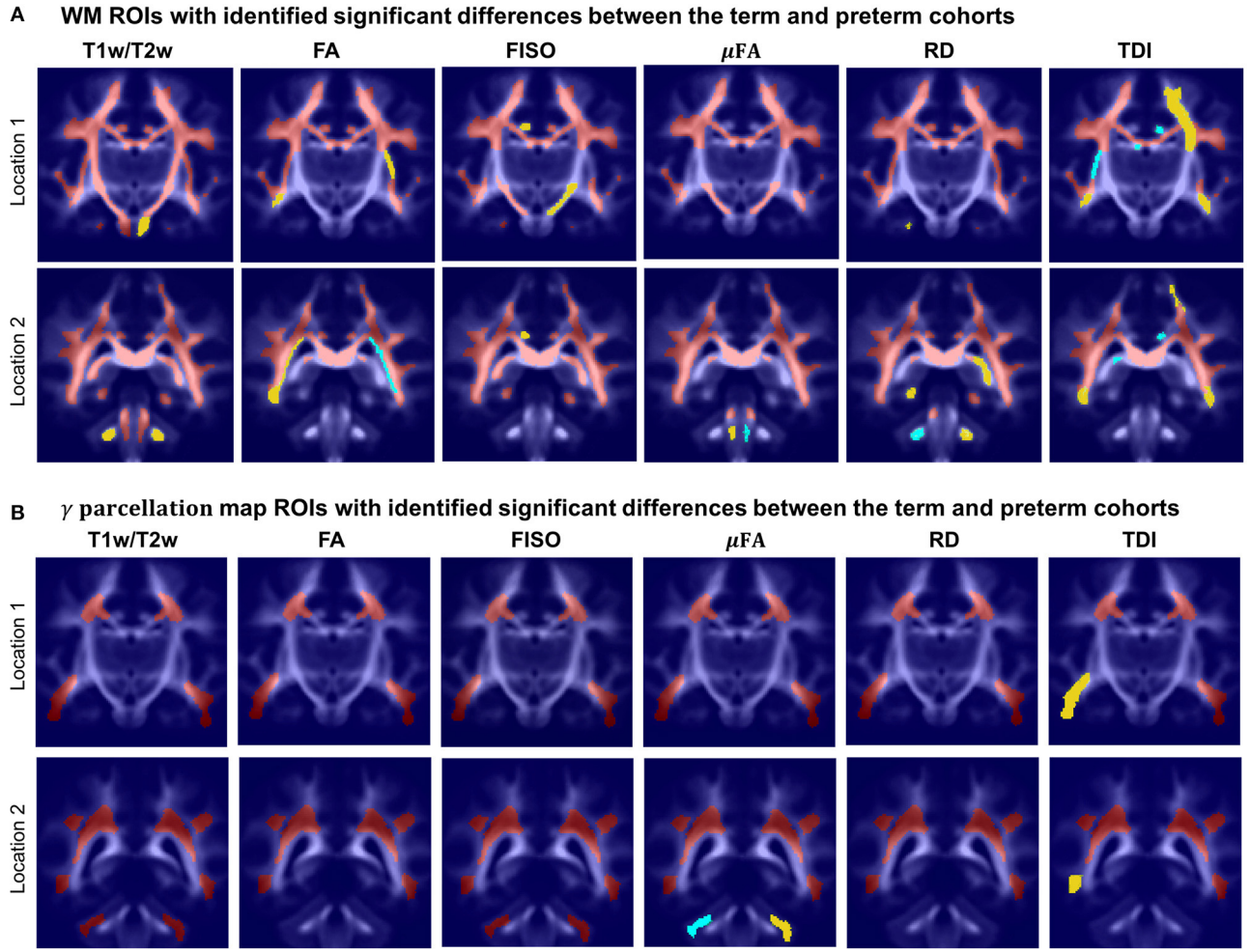


FIGURE 14 | Atlas-based region-specific analysis. The regions significantly associated with GA at birth are highlighted with red ($p < 0.001$), yellow ($p < 0.01$) and cyan ($p < 0.05$) and overlaid over the averaged TDI map in two coronal view locations. **(A)** WM parcellation regions. **(B)** γ^{av} parcellation regions.

3.4. Atlas-Based Region-Specific Analysis

In order to demonstrate the feasibility of the proposed MC atlas-based analysis approach and give an example of one of the possible applications, we performed ANOVA analysis to assess the influence of GA at birth on microstructure of WM regions delineated in our new atlas, with PMA as a confounding variable. To assess the feasibility of using the ANOVA analysis for the investigated datasets, we performed linear fitting for each of the channels. The γ^c values showed high correlation with the linear slope maps with the average NCC for all channels in the whole brain ROI 0.90 ∓ 0.09 (without CSF).

This is in agreement with the appearance of the global trends in **Figures 10–13** as well as the other reported studies (Feng et al., 2019; Dimitrova et al., 2020; O'Muircheartaigh et al., 2020) and confirms that during the short period between 37 and 44 weeks PMA range a linear approximation can be considered to be acceptable for ANOVA-based studies.

Figure 14 visualizes WM and transient regions in selected channels where average signal value was significantly associated with GA at birth. The main regions that have significant correlation of multiple indices with GA include: the corona radiata, superior longitudinal fasciculus, corpus callosum and thalamic radiation. The T1w/T2w contrast also showed to have significant correlation with GA in the internal and external capsule ROIs (**Figure 14A**). There is also a significant difference between the cohorts within the majority of γ^{av} parcellation regions (**Figure 14B**), which is in agreement with the expected prolonged existence of transient compartments in preterm subjects (Kostović and Judaš, 2006).

Figure 15A highlights the differences in the maturation rate γ^c maps between the term and preterm cohorts. The graphs in **Figure 15B** show the average signal values in the frontal right Von Monakow WM segment (highlighted in yellow in the γ^c maps). The rather wide range of values in all indices is potentially related to both the large size of the investigated WM region (approximately 3000 voxels) as well the individual variability also commonly observed in other neonatal brain studies (Feng et al., 2019; O'Muircheartaigh et al., 2020). There

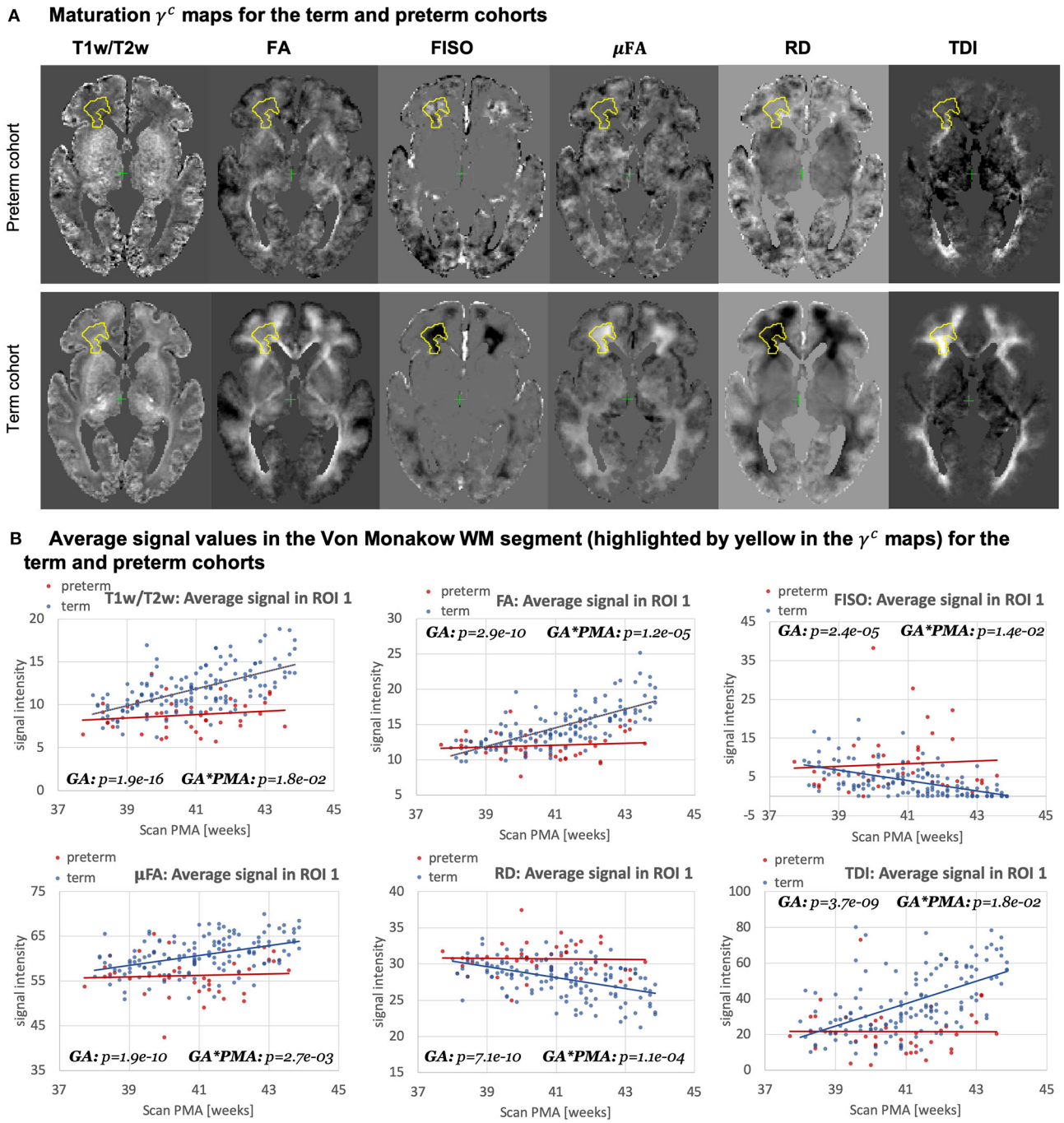


FIGURE 15 | Atlas-based analysis: comparison of the term (140) and preterm cohorts (40) for 38 to 43 weeks scan PMA range for a subset of channels c={ T1w/T2w; TDI; RD; FA; FISO; μFA }. **(A)** The γ^c maps of GF fitting for the term and preterm cohorts for 38 to 43 weeks PMA range. **(B)** The mean signal values in the frontal WM ROI from the $\gamma^{average}$ parcellation map (highlighted in yellow in the gamma maps) for the term (blue) and preterm (red) cohorts for 38–43 weeks PMA range.

is a clear increasing trend in T1w/T2w, FA and TDI for the term cohort along with decreasing FISO and RD. However, the slopes for the preterm cohort are close to zero with high variance in the signal values. Furthermore, in this region, the preterm subjects are characterized by significantly higher FISO and RD values and lower T1w/T2w, TDI and FA than the term cohort at the 42–43 week PMA period. This is consistent with the commonly reported lower FA and higher diffusivity values in preterm groups (Hermoye et al., 2006; Knight et al., 2018; Dimitrova et al., 2020), again suggesting delayed maturation of transient compartments in premature babies (Kostović and Judaš, 2006).

4. LIMITATIONS AND FUTURE WORK

The generated atlas is specific to the dHCP acquisition protocols, which might limit its application in terms of comparison with datasets from other studies. However, the proposed tools can be applied to generate study- and acquisition-specific 4D MC atlases. We investigated a relatively narrow neonatal period, and extension to a wider age range would improve the reliability of the Gompertz function fit and bring more insights into early brain development. In addition, a detailed region-specific statistical evaluation of the expected signal distributions of MRI-derived indices within the normal term cohort would need to be performed to allow accurate detection of image artifacts and brain abnormalities. Furthermore, the current work did not investigate the optimal preprocessing parameters required for fitting NODDI and μFA dMRI models, the effect of filtering (e.g., Smith et al., 2015) on the tractography outputs or the impact of different registration settings (e.g., channel weighting).

The study comparing term and preterm brain development included only 40 preterm subjects and they were not grouped with respect to specific types of anomalies, which can be addressed in future as more datasets become available. Furthermore, this work did not evaluate the influence of multi-channel registration on the extracted values of different microstructural indices. The generated WM parcellation map also potentially requires additional verification with respect to the correct definition of individual WM regions. Including additional cortical and sub-cortical regions or fixel-based analysis (Raffelt et al., 2017) could also enrich the insights into normal and preterm microstructural brain development.

5. CONCLUSIONS

In this work, we proposed and implemented a novel pipeline for generation of continuous 4D multi-channel atlases. It is based on multi-channel ODF+T2w+M_{cortex} guided registration and the Gompertz function fitting of both signal intensities and spatial transformations. The multi-channel registration pipeline implemented in MRtrix3 employs the novel local angular correlation similarity metric for ODF channels, LNCC metric for structural T2w and weighted fusion of the updates to the displacement fields. It also includes the cortex mask channel guided by LNCC metric for better alignment of the cortical regions.

Based on the proposed methods, we generated the first continuous multi-channel atlas of the normal term neonatal brain development during 37–44 weeks PMA generated from 170 subjects from the dHCP project. The atlas contains 14 channels including structural (T1w, T2w and T1w/T2w contrast) and DWI-derived metrics based on ODF, DTI, DKI, μFA and NODDI models. The Gompertz function fitting of the signal intensity and spatial transformation components in 4D allowed parametrization of the atlas. The output γ maps representing the rate of change can be used for interpretation of how maturation processes are manifested in different structural and diffusion MRI-derived metrics. Visual inspection of the fitting results showed that γ^c maps of the T2w, T1w, FISO, MD, RD and TDI channels are characterized by the high contrast in the fetal transient compartments (Pittet et al., 2019).

The atlas also includes two detailed WM parcellation maps: (i) the map with the major WM tract ROIs based on the definitions from the recently introduced M-CRIB-WM neonatal atlas (Alexander et al., 2020) and (ii) the map of the regions associated with high γ signal change rates during the normal WM maturation process. We tested the applicability of these parcellation maps for region-specific atlas-based studies on comparisons between the term and preterm cohorts. The results of this study showed significant effects linked to prematurity in multiple WM regions including the transient fetal compartments. The atlas and the software tools will be publicly available after publication of the article to support future studies of early brain development[1].

In summary, the proposed multi-channel registration and atlas facilitate combined analysis of structural and diffusion MRI indices in the same reference space without a bias from single-channel registration. Furthermore, combination of high resolution T2w and cortex mask channels with low resolution ODF channels aids better combined alignment of cortical and WM structures. To our knowledge, this is the first work that defines the pipeline for merged structural and diffusion MRI atlas-based analysis in neonatal brain studies.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions. The neonatal MRI datasets will be available in the dHCP project data release. Requests to access these datasets should be directed to http://www.developingconnectome.org/second-data-release.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by London - Riverside Research Ethics Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AU prepared the manuscript, implemented the code for the extended MC registration, fitting and analysis, generated the 4D MC atlas and conducted the experiments. IG participated in implementation of the preprocessing and analysis code, the design of the study and interpretation of the results. MP developed the original code for SSD MC ODF registration in MRtrix3. DB performed preprocessing of the dHCP datasets. MP, DC, and J-DT developed the tools for preprocessing and analysis of HARDI dHCP datasets. JH, EH, J-DT, LC, and AP developed MRI acquisition protocols for the neonatal dHCP datasets. LC developed the tools for preprocessing of structural dHCP datasets. JVH, ADE, SC, and MR are coordinators of the dHCP project. MD conceptualized the study and the methods, obtained the funding and supervised all stages of the research and

preparation of the manuscript. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

FUNDING

This work was supported by the Academy of Medical Sciences Springboard Award [SBF004\1040], MRC strategic grant [MR/K006355/1], European Research Council under the European Union's Seventh Framework Programme [FP7/20072013]/ERC grant agreement no. 319456 dHCP project, the Wellcome/EPSRC Centre for Medical Engineering at Kings College London [WT 203148/Z/16/Z], the NIHR Clinical Research Facility [CRF] at Guy's and St Thomas' and by the National Institute for Health Research Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. DC was supported by the Flemish Research Foundation [FWO], fellowship no. [12ZV420N].

ACKNOWLEDGMENTS

We thank everyone who was involved in acquisition and analysis of the datasets at the Department of Perinatal Imaging and Health at Kings College London. We thank all participants and their families. The formalization of the proposed MC registration method was reprinted by permission from our previous work (Uus et al., 2020).

REFERENCES

Akazawa, K., Chang, L., Yamakawa, R., Hayama, S., Buchthal, S., Alicata, D., et al. (2016). Probabilistic maps of the white matter tracts with known associated functions on the neonatal brain atlas: application to evaluate longitudinal developmental trajectories in term-born and preterm-born infants. *Neuroimage* 128, 167–179. doi: 10.1016/j.neuroimage.2015.12.026

Alexander, B., Yang, J. Y. M., Yao, S. H. W., Wu, M. H., Chen, J., Kelly, C. E., et al. (2020). White matter extension of the Melbourne Children's Regional Infant Brain atlas: M-CRIB-WM. *Hum. Brain Mapp.* 41, 2317–2333. doi: 10.1002/hbm.24948

Anderson, A. W. (2005). Measurement of fiber orientation distributions using high angular resolution diffusion imaging. *MR Med.* 54, 1194–1206. doi: 10.1002/mrm.20667

Andersson, J. L. R., Skare, S., and Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* 20, 870–888. doi: 10.1016/S1053-8119(03)00336-7

Andersson, J. L. R. and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* 125, 1063–1078. doi: 10.1016/j.neuroimage.2015.10.019

Avants, B. B., Duda, J. T., Zhang, H., and Gee, J. C. (2007). "Multivariate normalization with symmetric diffeomorphisms for multivariate studies," in *MICCAI 2007*, Vol. 4791 (LNCS) (Brisbane, QLD: Springer), 359–366.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Ball, G., Aljabar, P., Nongena, P., Kennea, N., Gonzalez-Cinca, N., Falconer, S., et al. (2017). Multimodal image analysis of clinical influences on preterm brain development. *Ann. Neurol.* 82, 233–246. doi: 10.1002/ana.24995

Barnett, M. L., Tusor, N., Ball, G., Chew, A., Falconer, S., Aljabar, P., et al. (2018). Exploring the multiple-hit hypothesis of preterm white matter damage using diffusion MRI. *Neuroimage Clin.* 17, 596–606. doi: 10.1016/j.nicl.2017.11.017

Bastiani, M., Andersson, J. L., Cordero-Grande, L., Murgasova, M., Hutter, J., Price, A. N., et al. (2019). Automated processing pipeline for neonatal diffusion MRI in the developing human connectome project. *Neuroimage* 185, 750–763. doi: 10.1016/j.neuroimage.2018.05.064

Batalle, D., O'Muircheartaigh, J., Makropoulos, A., Kelly, C. J., Dimitrova, R., Hughes, E. J., et al. (2019). Different patterns of cortical maturation before and after 38 weeks gestational age demonstrated by diffusion MRI in vivo. *Neuroimage* 185, 764–775. doi: 10.1016/j.neuroimage.2018.05.046

Bozek, J., Makropoulos, A., Schuh, A., Fitzgibbon, S., Wright, R., Glasser, M. F., et al. (2018). Construction of a neonatal cortical surface atlas using multimodal surface matching in the developing human connectome project. *Neuroimage* 179, 11–29. doi: 10.1016/j.neuroimage.2018.06.018

Christiaens, D., Cordero-Grande, L., Pietsch, M., Hutter, J., Edwards, A. D., Deprez, M., et al. (2018). "Multi-shell SHARD reconstruction from scattered slice diffusion MRI data in the neonatal brain," in *ISMRM 2018* (Paris), 464.

Christiaens, D., Cordero-Grande, L., Pietsch, M., Hutter, J., Price, A. N., Hughes, E. J., et al. (2021). Scattered slice SHARD reconstruction for motion correction in multi-shell diffusion MRI. *Neuroimage* 225:117437. doi: 10.1016/j.neuroimage.2020.117437

Claessens, N. H., Breur, J. M., Groenendaal, F., Wösten-van Asperen, R. M., Stegeman, R., Haas, F., et al. (2019). Brain microstructural development in neonates with critical congenital heart disease: an atlas-based diffusion tensor imaging study. *Neuroimage Clin.* 21:101672. doi: 10.1016/j.nicl.2019.101672

Cordero-Grande, L., Hughes, E. J., Hutter, J., Price, A. N., and Hajnal, J. V. (2018). Three-dimensional motion corrected sensitivity encoding reconstruction for multi-shot multi-slice MRI: application to neonatal brain imaging. *Magn. Reson. Med.* 79, 1365–1376. doi: 10.1002/mrm.26796

Counsell, S. J., Maalouf, E. F., Fletcher, A. M., Duggan, P., Battin, M., Lewis, H. J., et al. (2002). MR imaging assessment of myelination in the very preterm brain. *Am. J. Neuroradiol.* 23, 872–881. Available online at: http://www.ajnr.org/content/23/5/872

Dimitrova, R., Pietsch, M., Christiaens, D., Ciarrusta, J., Wolfers, T., Batalle, D., et al. (2020). Heterogeneity in brain microstructural development following preterm birth. *Cereb. Cortex* 30, 4800–4810. doi: 10.1093/cercor/bhaa069

Dubner, S. E., Rose, J., Bruckert, L., Feldman, H. M., and Travis, K. E. (2020). Neonatal white matter tract microstructure and 2-year language outcomes after preterm birth. *Neuroimage Clin.* 28:102446. doi: 10.1016/j.nicl.2020.102446

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., et al. (2012). 3D Slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* 30, 1323–1341. doi: 10.1016/j.mri.2012.05.001

Fenchel, D., Dimitrova, R., Seidlitz, J., Robinson, E. C., Batalle, D., Hutter, J., et al. (2020). Development of microstructural and morphological cortical profiles in the neonatal brain. *Cereb. Cortex* 30, 5767–5779. doi: 10.1093/cercor/bhaa150

Feng, L., Li, H., Oishi, K., Mishra, V., Song, L., Peng, Q., et al. (2019). Age-specific gray and white matter DTI atlas for human brain at 33, 36 and 39 postmenstrual weeks. *Neuroimage* 185, 685–698. doi: 10.1016/j.neuroimage.2018.06.069

Forsberg, D., Rathi, Y., Bouix, S., Wassermann, D., Knutsson, H., and Westin, C. F. (2011). "Improving registration using multi-channel diffeomorphic demons combined with certainty maps," in *MBIA*, Vol. 7012 (Toronto, ON: LNCS), 19–26. doi: 10.1007/978-3-642-24446-9_3

Geng, X., Styner,M., Gupta, A., Shen, D., and HGilmore, J. (2012). "Multi-contrast diffusion tensor image registration with structural MRI," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* (Barcelona), 684–687. doi: 10.1109/ISBI.2012.6235640

Girard, N., Dory-Lautrec, P., Koob, M., and Dediu, A. (2012). MRI assessment of neonatal brain maturation. *Imaging Med.* 4, 613–632. doi: 10.2217/iim.12.59

Glasser, M. F., and Van Essen, D. C. (2011). Mapping human cortical areas *in vivo* based on myelin content as revealed by T1- and T2-weighted MRI. *J. Neurosci.* 31, 11597–11616. doi: 10.1523/JNEUROSCI.2180-11.2011

Gupta, V., Malandain, G., Ayache, N., and Pennec, X. (2015). "A framework for creating population specific multimodal brain atlas using clinical T1 and diffusion tensor images," in *Computational Diffusion MRI. Mathematics and Visualization*, eds A. Fuster, A. Ghosh, E. Kaden, Y. Rathi, and M. Reisert (Munich: Springer). doi: 10.1007/978-3-319-28588-7_9

Hermoye, L., Saint-Martin, C., Cosnard, G., Lee, S.-K., Kim, J., Nassogne, M.-C., et al. (2006). Pediatric diffusion tensor imaging: normal database and observation of the white matter maturation in early childhood. *Neuroimage* 29, 493–504. doi: 10.1016/j.neuroimage.2005.08.017

Hughes, E. J., Winchman, T., Padormo, F., Teixeira, R., Wurie, J., Sharma, M., et al. (2017). A dedicated neonatal brain imaging system. *Magn. Reson. Med.* 78, 794–804. doi: 10.1002/mrm.26462

Hutter, J., Tournier, J. D., Price, A. N., Cordero-Grande, L., Hughes, E. J., Malik, S., et al. (2018). Time-efficient and flexible design of optimized multishell HARDI diffusion. *Magn. Reson. Med.* 79, 1276–1292. doi: 10.1002/mrm.26765

Iida, K., Takashima, S., and Ueda, K. (1995). Immunohistochemical study of myelination and oligodendrocyte in infants with periventricular leukomalacia. *Pediatr. Neurol.* 13, 296–304. doi: 10.1016/0887-8994(95)00192-1

Irfanoglu, M. O., Nayak, A., Jenkins, J., Hutchinson, E. B., Sadeghi, N., Thomas, C. P., et al. (2016). DR-TAMAS: diffeomorphic registration for tensor accurate alignment of anatomical structures. *Neuroimage* 132, 439–454. doi: 10.1016/j.neuroimage.2016.02.066

Jeurissen, B., Leemans, A., Tournier, J. D., Jones, D. K., and Sijbers, J. (2013). Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging. *Hum. Brain Mapp.* 34, 2747–2766. doi: 10.1002/hbm.22099

Jeurissen, B., Tournier, J. D., Dhollander, T., Connelly, A., and Sijbers, J. (2014). Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. *Neuroimage* 103, 411–426. doi: 10.1016/j.neuroimage.2014.07.061

Judaš, M., Radoš, M., Jovanov-Miloševic, N., Hrabac, P., Štern-Padovan, R., and Kostovic, I. (2005). Structural, immunocytochemical, and MR imaging properties of periventricular crossroads of growing cortical pathways in preterm infants. *Am. J. Neuroradiol.* 26, 2671–2684.

Kaden, E., Kruggel, F., and Alexander, D. C. (2016). Quantitative mapping of the per-axon diffusion coefficients in brain white matter. *Magn. Reson. Med.* 75, 1752–1763. doi: 10.1002/mrm.25734

Kellner, E., Dhital, B., Kiselev, V. G., and Reisert, M. (2016). Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magn. Reson. Med.* 76, 1574–1581. doi: 10.1002/mrm.26054

Kersbergen, K. J., Leemans, A., Groenendaal, F., van der Aa, N. E., Viergever, M. A., de Vries, L. S., et al. (2014). Microstructural brain development between 30 and 40 weeks corrected age in a longitudinal cohort of extremely preterm infants. *Neuroimage* 103, 214–224. doi: 10.1016/j.neuroimage.2014.09.039

Kimpton, J. A., Batalle, D., Barnett, M. L., Hughes, E. J., Chew, A. T., Falconer, S., et al. (2020). Diffusion magnetic resonance imaging assessment of regional white matter maturation in preterm neonates. *Neuroradiology* 63, 573–583. doi: 10.1007/s00234-020-02584-9

Knight, M. J., Smith-Collins, A., Newell, S., Denbow, M., and Kauppinen, R. A. (2018). Cerebral white matter maturation patterns in preterm infants: an MRI T2 relaxation anisotropy and diffusion tensor imaging study. *J. Neuroimaging* 28, 86–94. doi: 10.1111/jon.12486

Kostović, I., and Judaš, M. (2006). Prolonged coexistence of transient and permanent circuitry elements in the developing cerebral cortex of fetuses and preterm infants. *Dev. Med. Child Neurol.* 48, 388–393. doi: 10.1017/S0012162206000831

Krishnan, M. L., Wang, Z., Silver, M., Boardman, J. P., Ball, G., Counsell, S. J., et al. (2016). Possible relationship between common genetic variation and white matter development in a pilot study of preterm infants. *Brain Behav.* 6, 1–14. doi: 10.1002/brb3.434

Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A., et al. (2011). A dynamic 4D probabilistic atlas of the developing brain. *Neuroimage* 54, 2750–2763. doi: 10.1016/j.neuroimage.2010.10.019

Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M. A., Hajnal, J. V., and Schnabel, J. A. (2012). Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Med. Image Anal.* 16, 1550–1564. doi: 10.1016/j.media.2012.07.004

Kulikova, S., Hertz-Pannier, L., Dehaene-Lambertz, G., Buzmakov, A., Poupon, C., and Dubois, J. (2015). Multi-parametric evaluation of the white matter maturation. *Brain Struct. Funct.* 220, 3657–3672. doi: 10.1007/s00429-014-0881-y

Kunz, N., Zhang, H., Vasung, L., O'Brien, K. R., Assaf, Y., Lazeyras, F., et al. (2014). Assessing white matter microstructure of the newborn with multi-shell diffusion MRI and biophysical compartment models. *Neuroimage* 96, 288–299. doi: 10.1016/j.neuroimage.2014.03.057

Makropoulos, A., Aljabar, P., Wright, R., Hüning, B., Merchant, N., Arichi, T., et al. (2016). Regional growth and atlasing of the developing human brain. *Neuroimage* 125, 456–478. doi: 10.1016/j.neuroimage.2015.10.047

Makropoulos, A., Gousias, I. S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., et al. (2014). Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Trans. Med. Imaging* 33, 1818–1831. doi: 10.1109/TMI.2014.2322280

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., et al. (2018). The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage* 173, 88–112. doi: 10.1016/j.neuroimage.2018.01.054

Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol. Psychiatry* 80, 552–561. doi: 10.1016/j.biopsych.2015.12.023

Morel, B., Bertault, P., Favrais, G., Tavernier, E., Tosello, B., Bednarek, N., et al. (2021). Automated brain MRI metrics in the EPIRMEX cohort of preterm newborns: correlation with the neurodevelopmental outcome at 2 years. *Diagn. Interv. Imaging* 102, 225–232. doi: 10.1016/j.diii.2020.10.009

Oishi, K., Chang, L., and Huang, H. (2019). Baby brain atlases. *NeuroImage* 185, 865–880. doi: 10.1016/j.neuroimage.2018.04.003

Oishi, K., Mori, S., Donohue, P. K., Ernst, T., Anderson, L., Buchthal, S., et al. (2011). Multi-contrast human neonatal brain atlas: application to normal neonate development analysis. *Neuroimage* 56, 8–20. doi: 10.1016/j.neuroimage.2011.01.051

O'Muircheartaigh, J., Robinson, E. C., Pietsch, M., Wolfers, T., Aljabar, P., Grande, L. C., et al. (2020). Modelling brain development to detect white matter injury in term and preterm born neonates. *Brain* 143, 467–479. doi: 10.1093/brain/awz412

Pannek, K., Fripp, J., George, J. M., Fiori, S., Colditz, P. B., Boyd, R. N., et al. (2018). Fixel-based analysis reveals alterations is brain microstructure and macrostructure of preterm-born infants at term equivalent age. *Neuroimage Clin.* 18, 51–59. doi: 10.1016/j.nicl.2018.01.003

Pannek, K., Guzzetta, A., Colditz, P. B., and Rose, S. E. (2012). Diffusion MRI of the neonate brain: acquisition, processing and analysis techniques. *Pediatr. Radiol.* 42, 1169–1182. doi: 10.1007/s00247-012-2427-x

Pannek, K., Hatzigeorgiou, X., Colditz, P. B., and Rose, S. (2013). Assessment of structural connectivity in the preterm brain at term equivalent age using diffusion MRI and T2 relaxometry: a network-based analysis. *PLoS ONE* 8:e68593. doi: 10.1371/journal.pone.0068593

Park, H.-J., Kubicki, M., Shenton, M. E., Guimond, A., McCarley, R. W., Maier, S. E., et al. (2003). Spatial normalization of diffusion tensor MRI using multiple channels. *Neuroimage* 20, 1195–2009. doi: 10.1016/j.neuroimage.2003.08.008

Pecheva, D., Tournier, J. D., Pietsch, M., Christiaens, D., Batalle, D., Alexander, D. C., et al. (2019). Fixel-based analysis of the preterm brain: disentangling bundle-specific white matter microstructural and macrostructural changes in relation to clinical risk factors. *Neuroimage Clin.* 23:101820. doi: 10.1016/j.nicl.2019.101820

Pecheva, D., Yushkevich, P., Batalle, D., Hughes, E., Aljabar, P., Wurie, J., et al. (2017). A tract-specific approach to assessing white matter in preterm infants. *Neuroimage* 157, 675–694. doi: 10.1016/j.neuroimage.2017.04.057

Pietsch, M., Christiaens, D., Hutter, J., Cordero-Grande, L., Price, A. N., Hughes, E., et al. (2019). A framework for multi-component analysis of diffusion MRI data over the neonatal period. *Neuroimage* 186, 321–337. doi: 10.1016/j.neuroimage.2018.10.060

Pietsch, M., Raffelt, D., Dhollander, T., and Tournier, J.-D. (2017). "Multi-contrast diffeomorphic non-linear registration of orientation density functions," in *ISMRM 2017* (Honolulu, HI).

Pittet, M. P., Vasung, L., Huppi, P. S., and Merlini, L. (2019). Newborns and preterm infants at term equivalent age: a semi-quantitative assessment of cerebral maturity. *Neuroimage Clin.* 24:102014. doi: 10.1016/j.nicl.2019.102014

Raffelt, D., Tournier, J.-D., Fripp, J., Crozier, S., Connelly, A., and Salvado, O. (2011). Symmetric diffeomorphic registration of fiber orientation distributions. *Neuroimage* 56, 1171–1180. doi: 10.1016/j.neuroimage.2011.02.014

Raffelt, D., Tournier, J. D., Crozier, S., Connelly, A., and Salvado, O. (2012). Reorientation of fiber orientation distributions using apodized point spread functions. *MR Med.* 67, 844–855. doi: 10.1002/mrm.23058

Raffelt, D. A., Tournier, J. D., Smith, R. E., Vaughan, D. N., Jackson, G., Ridgway, G. R., et al. (2017). Investigating white matter fiber density and morphology using fixel-based analysis. *Neuroimage* 144, 58–73. doi: 10.1016/j.neuroimage.2016.09.029

Rose, J., Vassar, R., Cahill-Rowley, K., Stecher Guzman, X., Hintz, S. R., Stevenson, D. K., et al. (2014). Neonatal physiological correlates of near-term brain development on MRI and DTI in very-low-birth-weight preterm infants. *Neuroimage Clin.* 5, 169–177. doi: 10.1016/j.nicl.2014.05.013

Roura, E., Schneider, T., Modat, M., Daga, P., Muhlert, N., Chard, D., et al. (2015). Multi-channel registration of fractional anisotropy and T1-weighted images in the presence of atrophy: application to multiple sclerosis. *Funct. Neurol.* 30, 245–256. doi: 10.11138/FNeur/2015.30.4.245

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC.

Rutherford, M., Biarge, M. M., Allsop, J., Counsell, S., and Cowan, F. (2010). MRI of perinatal brain injury. *Pediatr. Radiol.* 40, 819–833. doi: 10.1007/s00247-010-1620-z

Schuh, A., Deprez, M., Makropoulos, A., Ledig, C., Counsell, S., V Hajnal, J., et al. (2014). "Construction of a 4D brain Atlas and growth model using diffeomorphic registration," in *Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data*, Vol. 8682 (Boston, MA), 27–37.

Schuh, A., Makropoulos, A., Robinson, E. C., Cordero-Grande, L., Hughes, E., Hutter, J., et al. (2018). Unbiased construction of a temporally consistent morphological atlas of neonatal brain development. *bioRxiv*. doi: 10.1101/251512

Schwartz, E., Kasprian, G., Jakab, A., Prayer, D., Schöpf, V., and Langs, G. (2016). "Modeling fetal cortical expansion using graph-regularized Gompertz models," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, Vol. 9900 (Athens), 247–254. doi: 10.1007/978-3-319-46720-7_29

Serag, A., Aljabar, P., Ball, G., Counsell, S. J., Boardman, J. P., Rutherford, M. A., et al. (2012). Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage* 59, 2255–2265. doi: 10.1016/j.neuroimage.2011.09.062

Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2015). SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *Neuroimage* 119, 338–351. doi: 10.1016/j.neuroimage.2015.06.092

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* 31, 1487–1505. doi: 10.1016/j.neuroimage.2006.02.024

Thompson, D. K., Kelly, C. E., Chen, J., Beare, R., Alexander, B., Seal, M. L., et al. (2019). Characterisation of brain volume and microstructure at term-equivalent age in infants born across the gestational age spectrum. *Neuroimage Clin.* 21:101630. doi: 10.1016/j.nicl.2018.101630

Tournier, J.-D., Calamante, F., and Connelly, A. (2010). "Improved probabilistic streamlines tractography by 2nd order integration over fiber orientation distributions," in *ISMRM 2010*, Vol. 1670 (Stockholm).

Tournier, J.-D., Christiaens, D., Hutter, J., Price, A. N., Cordero-Grande, L., Hughes, E., et al. (2020). A data-driven approach to optimising the encoding for multi-shell diffusion MRI with application to neonatal imaging. *NMR Biomed.* 33:e4348. doi: 10.1101/661348

Tournier, J.-D., Smith, R. E., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., et al. (2019). MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage* 202, 116–137. doi: 10.1016/j.neuroimage.2019.116137

Tournier, J. D., Calamante, F., and Connelly, A. (2007). Robust determination of the fiber orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage* 35, 1459–1472. doi: 10.1016/j.neuroimage.2007.02.016

Tournier, J. D., Calamante, F., and Connelly, A. (2012). MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22, 53–66. doi: 10.1002/ima.22005

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4itk: improved n3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Uus, A., Pietsch, M., Grigorescu, I., Christiaens, D., Tournier, J.-D., Grande, L. C., et al. (2020). "Multi-channel registration for diffusion MRI: longitudinal analysis for the neonatal brain," in Špiclin, Ž,*Biomedical Image Registration*, eds J. McClelland, J. Kybic, and O. Goksel (Cham: Springer International Publishing), 111–121.

Veraart, J., Novikov, D. S., Christiaens, D., Ades-aron, B., Sijbers, J., and Fieremans, E. (2016). Denoising of diffusion MRI using random matrix theory. *Neuroimage* 142, 394–406. doi: 10.1016/j.neuroimage.2016.08.016

Wang, S., Ledig, C., Hajnal, J. V., Counsell, S. J., Schnabel, J. A., and Deprez, M. (2019). Quantitative assessment of myelination patterns in preterm neonates using T2-weighted MRI. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-019-49350-3

Wright, R., Kyriakopoulou, V., Ledig, C., Rutherford, M. A., Hajnal, J. V., Rueckert, D., et al. (2014). Automatic quantification of normal cortical folding patterns from fetal brain MRI. *Neuroimage* 91, 21–32. doi: 10.1016/j.neuroimage.2014.01.034

Wu, D., Chang, L., Akazawa, K., Oishi, K., Skranes, J., Ernst, T., et al. (2017). Mapping the critical gestational age at birth that alters brain development in preterm-born infants Using multi-modal MRI. *Neuroimage* 149, 33–43. doi: 10.1016/j.neuroimage.2017.01.046

Young, J. M., Vandewouw, M. M., Morgan, B. R., Smith, M. L., Sled, J. G., and Taylor, M. J. (2018). Altered white matter development in children born very preterm. *Brain Struct. Funct.* 223, 2129–2141. doi: 10.1007/s00429-018-1614-4

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: Practical *in vivo* neurite orientation dispersion and density imaging of the human brain. *Neuroimage* 61, 1000–1016. doi: 10.1016/j.neuroimage.2012.03.072

Zhang, Y., Shi, F., Yap, P. T., and Shen, D. (2016). Detail-preserving construction of neonatal brain atlases in space-frequency domain. *Hum. Brain Mapp.* 37, 2133–2150. doi: 10.1002/hbm.23160

Zollei, L., Jaimes, C., Saliba, E., Grant, P. E., and Yendiki, A. (2019). TRActs constrained by underLying INfant anatomy(TRACULInA): an automated probabilistic tractography tool with anatomical priors for use in the newborn brain. *Neuroimage*, 199, 1–17. doi: 10.1016/j.neuroimage.2019.05.051

Disclaimer: The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 16 June 2021
doi: 10.3389/fnins.2021.636268

Fast and High-Resolution Neonatal Brain MRI Through Super-Resolution Reconstruction From Acquisitions With Variable Slice Selection Direction

Yao Sui[1,2]*, *Onur Afacan*[1,2], *Ali Gholipour*[1,2] *and Simon K. Warfield*[1,2]

[1] *Computational Radiology Laboratory, Department of Radiology, Boston Children's Hospital, Boston, MA, United States,*
[2] *Harvard Medical School, Boston, MA, United States*

The brain of neonates is small in comparison to adults. Imaging at typical resolutions such as one cubic mm incurs more partial voluming artifacts in a neonate than in an adult. The interpretation and analysis of MRI of the neonatal brain benefit from a reduction in partial volume averaging that can be achieved with high spatial resolution. Unfortunately, direct acquisition of high spatial resolution MRI is slow, which increases the potential for motion artifact, and suffers from reduced signal-to-noise ratio. The purpose of this study is thus that using super-resolution reconstruction in conjunction with fast imaging protocols to construct neonatal brain MRI images at a suitable signal-to-noise ratio and with higher spatial resolution than can be practically obtained by direct Fourier encoding. We achieved high quality brain MRI at a spatial resolution of isotropic 0.4 mm with 6 min of imaging time, using super-resolution reconstruction from three short duration scans with variable directions of slice selection. Motion compensation was achieved by aligning the three short duration scans together. We applied this technique to 20 newborns and assessed the quality of the images we reconstructed. Experiments show that our approach to super-resolution reconstruction achieved considerable improvement in spatial resolution and signal-to-noise ratio, while, in parallel, substantially reduced scan times, as compared to direct high-resolution acquisitions. The experimental results demonstrate that our approach allowed for fast and high-quality neonatal brain MRI for both scientific research and clinical studies.

Keywords: neonatal brain MRI, super-resolution, image reconstruction, anisotropic acquisition, isotropic reconstruction, fast imaging, spatial resolution, high-resolution MRI

OPEN ACCESS

Edited by:
Meritxell Bach Cuadra, Lausanne University Hospital (CHUV/UNIL), Switzerland

Reviewed by:
Jana Hutter, King's College London, United Kingdom
Sebastien Tourbier, Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland

***Correspondence:**
Yao Sui
yao.sui@childrens.harvard.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *01 December 2020*
Accepted: *19 May 2021*
Published: *16 June 2021*

Citation:
Sui Y, Afacan O, Gholipour A and Warfield SK (2021) Fast and High-Resolution Neonatal Brain MRI Through Super-Resolution Reconstruction From Acquisitions With Variable Slice Selection Direction. Front. Neurosci. 15:636268. doi: 10.3389/fnins.2021.636268

1. INTRODUCTION

Magnetic resonance imaging (MRI), as a noninvasive neuroimaging method, has revolutionized our knowledge over the past 20 years in understanding the human brain. Imaging for neonates and infants enables studying brain developments and neurodevelopmental disorders from early stages, which is crucially important to both scientific research and clinical studies (Weisenfeld and Warfield, 2009; Giampietri et al., 2015; Mongerson et al., 2019; Tortora et al., 2019; Ding et al., 2020). However, it is challenging to precisely delineate the anatomical structures of the

brain of neonates due to the small size of brain tissues in comparison to adults (Dubois et al., 2020). The spatial resolution is thus a critical factor in neonatal brain MRI. The typically used spatial resolutions in current clinical MRI practices, such as 3D imaging at isotropic 1 mm and 2D imaging with 0.5 mm in-plane resolution and 2 mm slice thickness, unfortunately, incur more partial voluming artifacts in neonates than in adults.

The interpretation and analysis of MRI of the neonatal brain benefit from a reduction in partial volume effect by increasing spatial resolution (Makropoulos et al., 2018; Dubois et al., 2019). Unfortunately, direct high-resolution (HR) MRI acquisition is time consuming and costly, and suffers from reduced signal-to-noise ratio (SNR). The long MRI scan for high spatial resolution potentially causes motion artifacts (Afacan et al., 2016). It is more prominent to the neonates who cannot be sedated, e.g., in a scan for the purpose of scientific research where sedation is typically unavailable. Also, even in scans where sedation is enabled to avoid subject motion, e.g., in a clinical scan, the long MRI scan for high spatial resolution leads to a substantial reduction in SNR, which in turn, increases the difficulty in distinguishing the signal of interest from noise. The underlying principle, from the imaging physics perspective, is that the reduced voxel size raises a reduction in the amount of signal received by the individual voxels. Consequently, the acquisition of short duration is critically important to neonatal MRI. The limitations of direct HR acquisition, therefore, necessitate the development of the methods that allow for imaging for neonates at high spatial resolution and high SNR, while in parallel, with short scan duration.

Current methods address the above limitations with a number of techniques, including parallel imaging (Pruessmann et al., 1999; Griswold et al., 2002), shifting to the ultra high field (7T) MRI (Annink et al., 2020), and super-resolution reconstruction (SRR) (Plenge et al., 2012). Parallel imaging and 7T MRI rely on hardware and imaging platforms, such as high density phased array receive coils and appropriate pulse sequence modification. In contrast, SRR, as a post-acquisition processing method, is performed on the acquired data that is in general of low spatial resolution and high SNR. Therefore, SRR is not subject to these limitations in hardware and platforms.

SRR originated in Tsai and Huang (1984) and was used for improving the quality of natural images. Fiat (2001) introduced SRR to MRI. It was showed in Scheffler (2002), and Peled and Yeshurun (2002) that SRR is unable to enhance the in-plane resolution of a 2D MRI or the resolution of a true 3D acquisition due to the Fourier encoding scheme. Also, it was demonstrated in Greenspan et al. (2002) that SRR is effective to improve the through-plane resolution of acquisitions of 2D slice stacks since the slices are individually Fourier encoded. Consequently, current SRR methods are designed to reduce the slice thickness of 2D slice stacks. Combining multiple low-resolution (LR) scans with different orientations was leveraged in Shilling et al. (2009), and then this framework was extended in Poot et al. (2010) to perform SRR with arbitrary image orientations and translations. SRR was quantitatively assessed and experimentally demonstrated in Plenge et al. (2012) to allow for a trade-off between spatial resolution, SNR, and acquisition time. Extensive SRR methods have recently been developed to improve MRI quality with a various of techniques (Gholipour et al., 2010a,b, 2015; Rousseau et al., 2010; Murgasova et al., 2012; Scherrer et al., 2012, 2015; Van Reeth et al., 2012; Kainz et al., 2015; Dalca et al., 2019; Sui et al., 2019, 2020).

SRR algorithms can mainly be classified as either a learning-based or a model-based method. Learning-based SRR summarizes the patterns mapping between LR and HR images over HR training data sets. Deep learning-based SRR has recently gained significant interest (Chaudhari et al., 2018; Chen et al., 2018; Zhao et al., 2019; Cherukuri et al., 2020; Wang et al., 2020; Xue et al., 2020). However, these methods require a large number of HR MRI acquisitions as the training data sets to learn the SRR model. The quality of the training data sets directly determines the quality of SRR. As discussed above, however, it is practically challenging to acquire HR data sets. Therefore, model-based SRR is commonly used in practice. Model-based SRR relies on an MRI acquisition model, from which an inverse problem is derived. As SRR estimates the super-resolved slices from much fewer acquired slices, the inverse problem is severely ill-posed. Prior knowledge, also known as regularization, is typically incorporated to separate the optimal estimate from the infinitely many solutions to the inverse problem. State-of-the-art priors include total variation (TV) (Plenge et al., 2012; Shi et al., 2015; Tourbier et al., 2015), non-local mean (Manjón et al., 2010), and gradient guidance Sui et al. (2019, 2020).

In this work, we developed a methodology for SRR based on the gradient guidance regularization method (Sui et al., 2019). It allows for high spatial resolution MRI with high SNR, excellent contrast-to-noise ratio (CNR), and reduced scan time, in comparison to direct HR acquisition. We achieved high quality brain MRI at a spatial resolution of isotropic 0.4 mm with 6 min of imaging time, using SRR from three short duration scans with variable directions of slice selection. Motion compensation is achieved by aligning the three short duration scans together. Our technique is thus suitable for use in a setting where direct HR acquisition is impractical. We applied this technique to 20 newborns and assessed the quality of the images we reconstructed. Experiments show that our SRR approach achieved considerable improvement in spatial resolution and SNR, while, in parallel, substantially reduced scan time, as compared to direct HR acquisition. The experimental results demonstrate that our approach allows for fast and high-quality neonatal brain MRI for both scientific research and clinical studies.

The novelty of this work is four-fold: (1) We take advantage of undersampling which allows us to form three undersampled neonatal scans with reduced acquisition time; (2) We encode the HR k-space data with three rapid undersampled observations of the HR k-space data convolved with a spatially oriented low-pass filter (being oriented axial, coronal, and sagittal). The estimation of the HR image from the undersampled observations is formulated as a deconvolution reconstruction problem; (3) The deconvolution reconstruction benefits from priors on edge position, which are easy to obtain and accurate in our setting; and (4) We apply our technique to neonatal brain MRI and achieve high quality images with reduced acquisition time.

2. MATERIALS AND METHODS

The purpose of our approach is to construct neonatal brain MRI images at isotropic high spatial resolution and high SNR with reduced acquisition time for both scientific research and clinical studies. We develop an SRR technique that can reconstruct isotropic HR images from multiple anisotropic acquisitions with variable directions in slice selection. To assess our approach, we simulated an MPRAGE data set based on images at an ultra high resolution of isotropic 250 μm and acquired 60 T2 FSE images from 20 newborns on a Siemens 3T scanner. In this section, we present the theory and algorithm used in our approach, the detailed descriptions of our data sets, the criteria used in the assessments, and the experimental designs, respectively.

2.1. Neonatal MRI Acquisition Strategy

As SRR is effective in enhancing the through-plane resolution of 2D slice stacks, we acquire the images with large matrix size and thick slices. The large matrix size ensures the in-plane high resolution while the use of thick slices enables short scan duration and high SNR. However, the thicker the slices, the more severe the partial volume effect, and thus the more difficult the super-resolution. To this end, we acquire multiple LR images to facilitate SRR, where an increased number of slices are acquired. However, the total acquisition time is increased accordingly due to the increased number of scans. Fortunately, we can employ fast imaging techniques to accelerate the scans, such as fast spin echo (FSE) imaging. For images that yield long repetition time (TR), such as T_2-weighted images, the FSE technique can significantly reduce the scan duration by n_{ETL} times with an echo train length (ETL) of n_{ETL} that typically ranges from 4 to 32 in clinical routines.

The goal of SRR is to estimate the missing signal in k-space based on the sampled k-space data. Our approach performs the estimation in the spatial image domain, which relates to the k-space data through Fourier transforms. Variable slice selection directions are incorporated in the acquisitions of the LR scans, where each LR scan contains a certain amount of k-space data in the slice selection direction. Consequently, the LR scan set comprises the spatial frequencies in different directions in the 3D frequency spectrum space. By combining multiple such LR scans, the difficulty of the SRR is thus reduced as an increased amount of k-space data is sampled. Although the slice selection directions and the number of the LR scans can be arbitrary, orthogonal (axial, coronal, and sagittal) acquisitions typically achieved a trade-off between acquisition time and SRR performance, since the acquired data yields the three complementary imaging planes.

We acquire three T2 FSE images from each neonate with variable directions in slice selection, which are typically carried out in three complementary planes (axial, coronal, and sagittal), and perform SRR to form an isotropic HR image. We set the parameters according to the scan time:

$$T \simeq TR \cdot \left\lceil \frac{FoV_p}{S_p \cdot f_{acc} \cdot ETL} \right\rceil \cdot N_{NEX}, \tag{1}$$

where FoV_p denotes the Field of view (FoV) in the phase encoding direction, S_p denotes the voxel size in the phase encoding direction, f_{acc} is the acceleration factor of parallel imaging, ETL is the echo train length, N_{NEX} is the number of excitations, and $\lceil x \rceil$ returns the smallest integer that is $>x$. We recommend that FoV_p ranges from 120 to 150 mm to fit the head size of the subject. S_p is kept at 0.39 mm. GRAPPA parallel imaging is leveraged with an acceleration factor of 2. Averaging is not considered in our fast imaging protocol, so $N_{NEX} = 1$. We recommend using ETL between 16 and 21 for fast scans of high quality. TR is typically set over 10 s depending on the number of slices required as well as the head size of the subject. We typically acquire 60–80 slices per image, and the slice thickness is fixed at 2 mm. It takes <2 min with our fast imaging protocol to acquire a T2 FSE image, i.e., $T \leq 120$ s. For the largest value of FOV, i.e., $FoV_p = 150$ mm, with an ETL =21, it allows a TR ≤13.1 s according to (1), which is a sufficiently high value for TR. Consequently, our protocol can ensure less than two minutes of imaging time to acquire a T2 FSE image at the in-plane resolution of 0.39 mm for a neonate. Besides the parameters related to the scan time, we set TE = 93 ms, flip angle = 160°, and echo spacing = 9.8 ms. We use an interleaved acquisition mode, with which an even-first ascending slice order with an interleave factor of 2 is incorporated, i.e., the slice order is $[2:2:N, 1:2:N-1]$ for an image with N slices. The HR image is reconstructed at the resolution of isotropic 0.39 mm, which is sufficiently high for the interpretation and analysis of the anatomical structures of the neonatal brain in clinical practices.

2.2. Neonatal MRI Reconstruction

We leverage the gradient guidance regularized SRR algorithm (Sui et al., 2019) to reconstruct the neonatal MRI images[1]. Given n acquired LR images $\{\mathcal{Y}\}_{k=1}^{n}$, the forward model that describes the MRI acquisition process can be found from the HR image $\mathcal{X}$ by

$$\mathbf{y}_k = \mathbf{D}_k\mathbf{H}_k\mathbf{T}_k\mathbf{x} + \boldsymbol{\varepsilon}_k, \quad , k = 1, 2, 3, \ldots, n, \tag{2}$$

where $\mathbf{y}_k$ and $\mathbf{x}$ are column vector form of $\mathcal{Y}_k$ and $\mathcal{X}$, respectively; $\mathbf{T}_k$ denotes a coordinate transform of $\mathcal{X}$ in the 3D space; $\mathbf{H}_k$ denotes a blur kernel; $\mathbf{D}_k$ denotes a downsampling operation; and $\boldsymbol{\varepsilon}_k$ denotes the imaging noise.

The noise $\boldsymbol{\varepsilon}_k$ can be considered as additive and Gaussian when SNR>3 (Hansen and Kellman, 2015). Therefore, the noise in each acquisition can be independently formulated as an identical Gaussian distribution. The HR reconstruction $\mathbf{x}$ is consequently obtained by solving the inverse problem

$$\min_{\mathbf{x}} \sum_{k=1}^{n} \left\| \mathbf{D}_k\mathbf{H}_k\mathbf{T}_k\mathbf{x} - \mathbf{y}_k \right\|_2^2 + \lambda \sum_{s\in\mathcal{S}} \left\| \nabla_s\mathbf{x} - \mathbf{g}_s \right\|_1, \tag{3}$$

where $\mathcal{S}$ indexes a set of spatial image gradients, $\mathbf{g}_s$ denotes the s-th component of the gradient guidance, $\nabla_s\mathbf{x}$ computes the s-th spatial gradient of $\mathbf{x}$, which is calculated from the same orientation and the same scale as $\mathbf{g}_s$, and $\lambda > 0$ is a weight parameter for the regularization term. The above minimization can be accomplished by a subgradient descent (Bertsekas, 1999)

[1]The term "image" indicates a volumetric image of slice stack here and hereafter.

or a proximal gradient descent algorithm (Daubechies et al., 2003).

As the images are acquired fast, we consider that there is no intra-volume head motion during the acquisition. Therefore, the transform $\mathbf{T}_k$ in Equation (3) compensates for the misalignment between scans. We use a rigid body transform to represent the misalignment. Consequently, $\mathbf{T}_k$ is defined by the parameters of six degrees of freedom (three for rotation and three and translation). We first interpolate all the LR images to those of the same size and the same resolution as the HR image being reconstructed by using a third-order B-spline interpolation method. We set $\mathbf{T}_1$ to an identity transform and evaluate $\mathbf{T}_k$ for $k > 1$ by aligning the k-th interpolated LR image to the first interpolated LR image. In the alignments, mutual information is leveraged to measure the similarity between the first and k-th images. We use the CRKIT[2] to accomplish the image alignment.

The blur kernel $\mathbf{H}_k$ in Equation (3) is a spatial invariant operator. It raises the partial volume effect in the acquired image. As only the through-plane resolution is enhanced while the in-plane resolution is kept unchanged, we design the blur kernel as a low-pass filter in the slice selection direction, also known as the slice profile. In the MRI acquisition, each slice is excited by incorporating a selective gradient that is generated by the radio frequency (RF). Ideally, the slice profile is desired to be a boxcar function. This requires infinitely many frequencies to yield the RF, which are impossible to obtain in practice. It is crucial to appropriately approximate the slice profile in SRR as the approximation directly influences the accuracy of the forward model. In general, bell-curve profiles with wider bases and narrower central peaks are leveraged, and slice thickness is measured as the full width at half maximum (FWHM) signal intensity. Gaussian profiles are widely used in MRI reconstruction and have been demonstrated to be effective in SRR (Rousseau et al., 2005; Jiang et al., 2007; Gholipour et al., 2010a,b; Murgasova et al., 2012; Sui et al., 2019, 2020). Therefore, we approximate the slice profile by a Gaussian function with an FWHM equal to the slice thickness.

As the downsampling factor can be arbitrary, instead of an integer for natural images, it is inconvenient to perform the downsampling in the image domain. Consequently, the downsampling operator $\mathbf{D}_k$ in Equation (3) is implemented in the frequency domain by cropping out the low frequencies. The respective upsampling operation is thus implemented by inserting zeros at the missing high frequencies. In our implementations, we combine the Gaussian profile and the downsampling operator into a single filter in the frequency domain for computational efficiency. As the Gaussian profile is performed in a manner of a low-pass filter, truncating high frequencies for downsampling does not cause intensity oscillations in the image domain.

The spatial image gradient guides the HR reconstruction. The index set $\mathcal{S}$ in the regularization term of Equation (3) comprises 40 spatial gradient fields that yield different orientations and different scales, as suggested in Sui et al. (2019). All the 40 gradient fields are combined into a gradient guidance, denoted by $\mathbf{g}$ in Equation (3). The s-th component of $\mathbf{g}$ is separately computed from the image constructed by the interpolation and average (IAA) method. In the IAA method, the n aligned LR images are interpolated to the same size at the same resolution as the HR reconstruction, and then the reconstructed HR image is formed by averaging out the n interpolated images. Specifically, with an image obtained by IAA, denoted by I, a component of the gradient guidance is calculated by $I - D_x^\alpha D_y^\beta D_z^\gamma I$ where D_m^n denotes the operation that circularly shifts an image in $m \in \{x, y, z\}$ direction by n voxels. We set α to integers between -2 and 2, and β and γ between 0 and 2. We exclude the components calculated at $\alpha = \beta = \gamma = 0$ and $\alpha + \beta + \gamma < 0$ to eliminate the replicates. Consequently, we have 40 components calculated for the gradient guidance. We put all the 40 components in a set $\mathcal{S}$ and index them by $\mathbf{g}_s$ in Equation (3). We set the regularization weight parameter λ in Equation (3) to 0.1 in all experiments in this paper according to our experimental investigation.

The source codes and a docker version of our reconstruction algorithm can be checked out from our website[3].

2.3. Assessment Criteria

We assess our approach in terms of spatial resolution, SNR, CNR, and acquisition time.

2.3.1. Spatial Resolution

The signal intensity in a voxel is quantified as the integration of the signal over a spatial region defined by the position and size of the voxel. Spatial resolution is usually used to describe in an image the number of *independent* voxels per unit length or volume. Different from the measure based on voxel size, spatial resolution refers to the ability to differentiate two types of brain tissues that are relatively close together. As partial voluming artifacts occur due to the dependent voxels, the number of voxels suffering from partial volume effects can be an effective measure for spatial resolution. The higher the spatial resolution, the fewer the voxels affected by partial volume effects. To this end, we evaluate the percentage of the voxels that comprise the signal from more than one type of brain tissue and use it as the metric of the partial volume effect estimation. An image at a higher spatial resolution thus yields a lower metric value of partial volume effect.

We consider three types of brain tissues in the estimation of the partial volume effect from the neonatal MRI reconstruction: cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM). As the three tissues yield different contrasts in MR images, the intensities of the voxels from them scatter in three clusters. Due to the partial voluming, there may be overlaps in the three clusters. We thus investigate the distribution of the voxel intensities of the HR reconstruction. First, we select an image region that contains the three tissues, and then construct a histogram of the voxel intensities over the selected image region. It has been shown in Laidlaw et al. (1998) that the distribution of voxel intensities from a pure tissue is Gaussian. Therefore, we fit the histogram of the voxel intensities by a Gaussian mixture

[2]CRKIT - Computational Radiology Kit, http://crl.med.harvard.edu/software/.

[3]We are preparing the codes and the docker file, and will make them publicly available after the paper is accepted.

model (GMM) with three components that characterize the three types of brain tissues. The voxels from a pure tissue are thus identified if their intensities range from $\mu - \delta$ to $\mu + \delta$ for μ and δ being the mean and half of FWHM of the corresponding Gaussian component in the GMM, respectively. We apply the GMM to the entire image to form the voxel set-1 containing all voxels from the three tissues (i.e., voxels may contain the signal from more than one tissues) and set-2 consisting of the voxels identified from each pure tissue (i.e., voxels contain the signal from only one tissue). The difference between the two sets of voxels consequently indicates the number of voxels suffering from the partial volume averaging.

2.3.2. SNR and CNR

We compute the SNR of an image from the mean of signal intensities over the noise. Specifically, the SNR is found by $SNR = 10\log_{10}\frac{\sum_{k=1}^{3} w_k s_k}{\sigma \sum_{k=1}^{3} w_k}$ where s_k and w_k denote the mean signal intensity of the voxels and the percentage of the voxels from the k-th pure tissue, respectively, and σ denotes noise measure. Both s_k and w_k can be directly obtained from the fitted GMM constructed above. s_k is computed from the mean of the k-th Gaussian component, while w_k is evaluated as the maximum of the k-th Gaussian component. We select an image region in the background and compute the standard deviation of the voxel intensities over the region as the noise measure.

Similar to SNR, we compute the CNR from the difference of the mean of signal intensities between two types of tissues over the noise: $CNR_{j,k} = 10\log_{10}\frac{|s_i - sj|}{\sigma}$ We evaluate in the assessment the CNR between CSF and GM, denoted by CNR:CSF-GM, the CNR between CSF and WM, denoted by CNR:CSF-WM, and the CNR between GM and WM, denoted by CNR:GM-WM.

2.4. Experimental Design

We conduct two experiments to assess our approach on simulated data as well as the data acquired from 20 newborns on a Siemens 3T scanner. The goal of the experiments is to demonstrate that our approach achieves high-quality neonatal brain MRI with reduced imaging time, which allows for the studies with both research and clinical purposes.

We leveraged two other acquisition strategies as baseline schemes to compare to in the experiments, including direct HR acquisition (DA) and the single image-based super-resolution (SISR) method. Our approach was assessed by comparing to DA to verify the improved image quality and reduced acquisition time. The SISR used the same SRR algorithm as our approach with the same parameters setting. It is in fact a special case of our approach when only one LR scan was acquired, i.e., $n = 1$ in Equation (3). We used approximately three times more slices in a single LR image than in an LR image in our approach, in order to ensure equal acquisition time (by conducting the same number of phase encoding steps) for a fair comparison. Consequently, the comparisons to SISR evaluated the superiority of our approach to variable slice selection direction over the acquisition with constant slice selection direction.

We employed other four state-of-the-art SRR methods as baseline methods to assess our approach, including the interpolation and average (IAA) method, total variation (TV) prior (Plenge et al., 2012), non-local upsampling (NLU) method (Manjón et al., 2010), and a deep convolutional network-based SRR (SRCNN) method (Dong et al., 2016). The IAA method interpolated the n LR images to the same size at the same resolution as the HR reconstruction by a third-order B-spline method, and then aligned all the interpolated images together. The reconstruction was finally obtained by averaging all the interpolated and aligned images. IAA is one of the most widely used methods in both clinical practices and scientific research studies due to its effectiveness in improving SNR. We therefore compared our approach to IAA to assess the applicability of our approach in practical imaging tasks. The NLU method further processed the results generated by IAA with a non-local mean algorithm. The TV method used the same deconvolution scheme as our approach to reconstruct the HR image. Our scan strategy allows for training deep 2D SRR models as it acquires in-plane HR slices. The deep SRR models can be trained on these HR slices and then used to super-resolve the through-plane LR slices to generate an isotropic HR image. Although recent years have witnessed the extensively proposed deep neural networks-based SRR methods, only lightweight deep architectures allow for the training due to the limited number of HR slices acquired with our scan strategy. We therefore employed SRCNN as a deep baseline model in the experiments, which comprises about 8 k parameters to train. The trained model was applied to the through-plane LR slices of each LR image. The reconstructed HR image was formed by averaging out all the super-resolved images on their voxels. We set the weight parameter of the TV method at 0.1 for its best results according to the simulation results. We set the parameters in NLU according to the recommendation in Manjón et al. (2010).

2.4.1. Experiment 1: Simulations on MPRAGE Data

The goal of this experiment is three-fold: (1) to investigate the influence of the gradient guided regularization on the SRR performance; (2) to demonstrate that our anisotropic acquisition strategy with variable directions in slice selection leads to superior SRR to the strategy of single acquisition; and (3) to show that our SRR approach achieves the MR images of higher quality than direct HR acquisition in terms of spatial resolution and SNR.

For the experimental goal, we simulated a data set based on the Dryad data set containing eight MPRAGE images at an ultra high resolution of isotropic 250 μm (Lusebrink et al., 2017). This data set was acquired from an adult subject, and the acquisition time was about 1 h per image with very complicated protocols and pre- and post-acquisition processing operations, in order to preserve a satisfactory SNR. So it is practically impossible to acquire such images in clinical routines. As there is currently no publicly available HR neonatal brain scan, and it is challenging to acquire an HR image from a neonate at a satisfactory SNR, we used this data set for the simulation demonstrations. Considering the goal of this experiment addressed above, it is reasonable to use this data set for the demonstrations.

We generated eight images at the resolution of isotropic 0.5 mm by downsampling each original image, and used them as the direct HR acquisitions. The downsampling followed the process

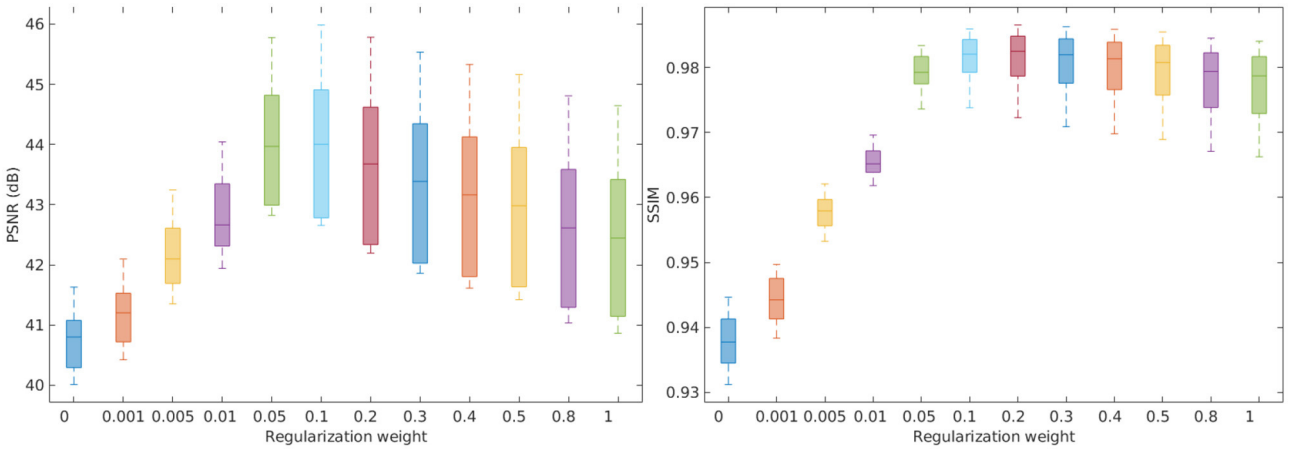


FIGURE 1 | Investigation on the influence of the gradient guided regularization to the SRR performance in terms of PSNR and SSIM. The results show that our SRR approach performed the best with the regularization weight parameter λ ranging from 0.05 to 0.3. The results also suggest that the regularization considerably improved the SRR performance on the simulation data set by referring to the results at $\lambda = 0$ (in the case of no regularization).

defined in the forward model, as shown in Equation (2). Then, we simulated three LR images based on each direct HR acquisition in the three complementary planes and used them as our anisotropic acquisitions with variable directions in slice selection. The in-plane resolution of these LR images was 0.5 × 0.5 mm and the slice thickness was 2 mm. To keep the contrast unchanged, we assumed the echo time (TE) and repetition time (TR) of these LR images the same as the direct HR acquisitions. Each direct HR acquisition comprised 193,600 phase encoding steps, while the three LR images contained 132,000 phase encoding steps in total. Therefore, the acquisition time of the three LR images was ∼68% of that of the direct HR acquisition. For the single acquisition-based SRR, we generated an LR image at the resolution of 0.5 × 0.5 × 0.73 mm. The resolution was derived from that the same number of phase encoding steps (132,000 steps) were conducted for this image. All the simulations for the LR images followed the process described in Equation (2).

We investigated the regularization weight parameter λ in Equation (3) to study the influence of the gradient guided regularization on the SRR performance. We ran our SRR approach with different λ values in a certain range and evaluated the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) (Wang et al., 2004) against the ground truth image in the simulation experiment. We fixed the value of λ in all other experiments reported in this paper according to the investigation results.

We reconstructed the HR images at the resolution of isotropic 0.5 mm by using the gradient guidance regularized SRR algorithm, as shown in Equation (3) on the data sets simulated from the anisotropic and single acquisition strategies, respectively. We compared the HR images reconstructed by our approach to the HR reconstructions by SISR and the direct HR acquisitions in terms of the spatial resolution, SNR, and CNR. Through the comparisons, we can answer the questions: (1) can SRR constructs images of higher quality with lower acquisition time than direct HR acquisition? and (2) with the same acquisition time, which acquisition strategy leads to better SRR, our anisotropic acquisition or the single acquisition? The second question is essentially about how we allocate data acquisitions for a better SRR given a fixed acquisition time.

2.4.2. Experiment 2: Assessment on Clinical T2 FSE Data

The objective of this experiment is to evaluate our approach on the clinical data and to demonstrate that our approach can provide high quality images for both scientific research studies and clinical routines in neonatal brains. To this end, we acquired a data set with the protocol presented above. The data set comprised 60 neonatal brain MR images acquired from 20 newborns (acquired three from each). All scans were performed in accordance with the local institutional review board (IRB) protocol. We incorporated the IAA method as a baseline in this experiment, which is one of the most widely used methods in both clinical practices and scientific research studies. The HR images were reconstructed at the resolution of isotropic 0.39 mm in this experiment by our approach and the IAA method. These reconstructed HR images were assessed in terms of spatial resolution, SNR, and CNR.

3. RESULTS

Our reconstruction algorithm was implemented in MATLAB (The MathWorks Inc.) without any code optimizations. We carried out our algorithm on a workstation with an Intel Xeon CPU@2.1 GHz and 128 GB memory. It took about 15 min to reconstruct an image of the typical size 384 × 384 × 384 voxels.

We reported and visualized our quantitative results by using the box and whisker plot (McGill et al., 1978; Langford, 2006). On each box, the central mark indicated the median, and the bottom and top edges of the box indicated the 25th and 75th

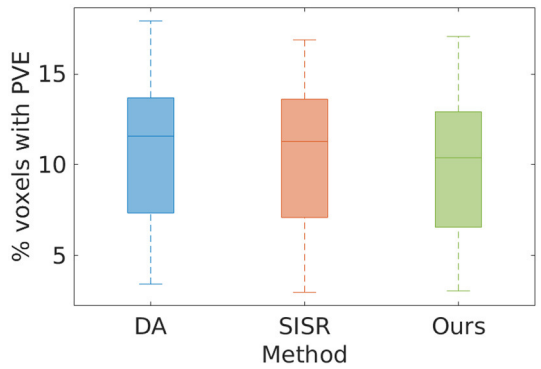


FIGURE 2 | Estimates of partial volume effects (PVE) from the eight MPRAGE images directly acquired and reconstructed by SISR and our approach on the simulated data set, respectively. The average percentages of the voxels suffering from PVE were, respectively 10.82 ± 5.02, 10.48 ± 4.93, and 9.99 ± 4.80% with the methods of direct acquisition (DA), SISR, and ours. The results show that our approach generated the highest spatial resolution on this data set. Our approach yielded a 7.7% reduction in the partial volume effects, leading to the enhancement in spatial resolution, as compared to the direct HR acquisitions. The results also suggest that SRR (both SISR and our approach) achieved higher spatial resolution with much lower acquisition time than direct HR acquisition.

percentiles, respectively. The whiskers extended to the most extreme data points.

3.1. Experiment 1: Simulations on MPRAGE Data

Figure 1 shows the investigation results on the influence of the gradient guided regularization on the SRR performance in terms of PSNR and SSIM. The results show that our SRR approach performed the best with the regularization weight parameter λ ranging from 0.05 to 0.3. The results also suggest that the regularization considerably improved the SRR performance on the simulation data set by referring to the results at $\lambda = 0$ (in the case of no regularization). According to the investigation results, we therefore fixed the regularization weight parameter λ at 0.1 in all the rest experiments reported in this paper.

Figure 2 shows the estimates of partial volume effect from the eight MPRAGE images directly acquired and reconstructed by SISR and our approach on the simulated data set, respectively. The average percentages of the voxels suffering from partial volume effect were respectively 10.82 ± 5.02%, 10.48 ± 4.93%, and 9.99 ± 4.80% with the methods of direct acquisitions, SISR, and ours. The results show that our approach generated the highest spatial resolution on this data set. Our approach yielded a 7.7% reduction in the partial volume effects, leading to the enhancement in spatial resolution, as compared to the direct acquisitions the resolution of isotropic 0.5 mm. The results also suggest that SRR (both SISR and our approach) achieved higher spatial resolution with much lower acquisition time than direct HR acquisition.

Figure 3 shows the results of direct acquisition, SISR, and our approach in terms of SNR and CNR from the eight MPRAGE acquisitions/reconstructions on the simulated data set. The average SNRs obtained from direct acquisition, SISR, and our approach were 14.51 ± 0.57, 15.66 ± 0.48, and 16.62 ± 0.56 dB, respectively. Our approach achieved higher SNR on this data set, and yielded 2.11 dB enhancement in SNR as compared to the direct acquisitions. Two-sample t-test at the 5% significance level showed that our approach significantly outperformed DA ($p = 3.02e^{-6}$) and SISR ($p = 2.40e^{-3}$). Wilcoxon signed-rank tests, where the null hypothesis was the difference of two sets of data comes from a distribution with zero median, showed that the population mean rank of our approach significantly differed from the two baselines in SNR at the 5% significance level (rejected the null hypothesis with $p = 7.8e^{-3}$ for both DA and SISR). Our approach consistently offered the highest CNRs between the three types of brain tissues on this data set. In particular, our approach achieved 1.31 dB higher CNR between GM and WM than direct acquisition. The results show that SRR led to considerably improved SNR and CNR as compared to direct HR acquisition.

Figure 4 shows the qualitative results in representative slices from the images directly acquired, reconstructed by SISR and our approach, respectively. The slices directly acquired and formed by SISR were much noisy as compared to our reconstructions. The noise was more prominent for SISR in the voxels from the skull, as highlighted by the red arrows. Although what we were interested in were CSF, GM, and WM, the noisy voxels from the skull rendered that SISR generated noise all over the images but just not as obvious as those from the skull.

3.2. Experiment 2: Assessment on Clinical T2 FSE Data

Figure 5 shows the quality of the 20 HR images reconstructed by the five SRR methods on the clinical data set in terms of SNR and CNR. The average SNR achieved by the five methods are, respectively: IAA = 20.19 ± 2.57 dB, TV = 19.17 ± 3.40 dB, NLU = 19.92 ± 2.04 dB, SRCNN = 20.18 ± 1.98 dB, Ours = 20.04 ± 2.77 dB. IAA, NLU, and SRCNN generated high SNR, as they benefited from the averaging to improve the SNR and CNR. Our approach offered comparable SNR with IAA, NLU, and SRCNN, and outperformed TV by ∼1 dB in terms of SNR. Our approach generated slightly superior CNRs to the five baselines about cerebrospinal fluid, and yielded considerably higher CNR between gray matter and white matter than these baselines.

Figure 6 the spatial resolution evaluated from the twenty images reconstructed by the four baselines and our approach on the clinical data set in terms of partial volume effect. The average PVE achieved by the five methods are, respectively: IAA = 19.40 ± 11.85%, TV = 9.02 ± 7.30%, NLU = 11.35 ± 7.69%, SRCNN = 10.88 ± 7.46%, Ours = 7.25 ± 4.37%. Our approach offered a considerably lower percentage of the voxels suffering from partial volume averaging in the HR reconstructions than the four baselines, leading to substantially enhanced spatial resolution. Two-sample t-test at the 5% significance level showed that our approach significantly outperformed IAA ($p = 1.40e^{-6}$), NLU ($p = 2.65e^{-4}$), and SRCNN ($p = 5.39e^{-4}$). Wilcoxon signed-rank tests, where the null hypothesis was the difference of two sets

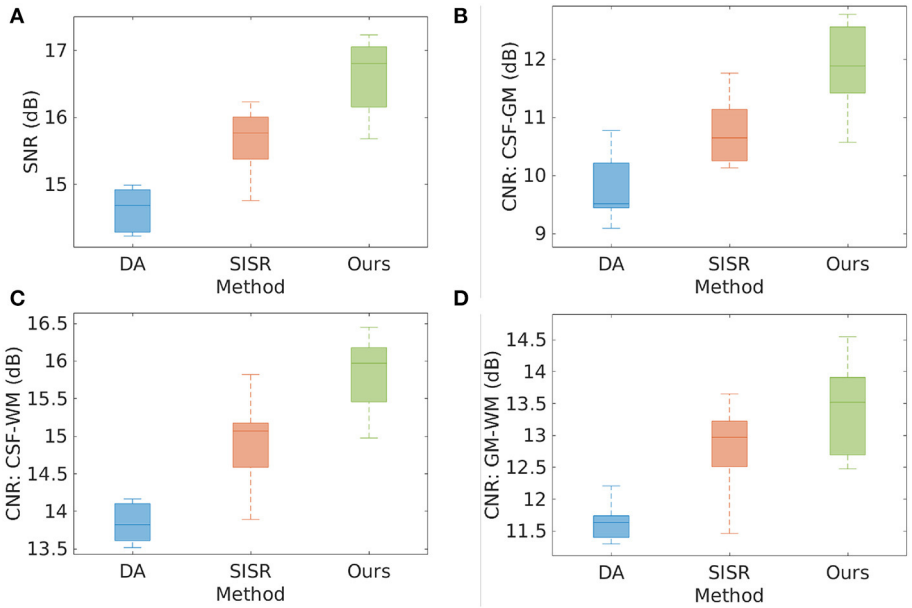


FIGURE 3 | Results of direct acquisition (DA), SISR, and our approach in terms of SNR and CNR from the eight MPRAGE acquisitions/reconstructions from the simulated data set. **(A)** The average SNRs obtained from DA, SISR, and our approach were 14.51 ± 0.57, 15.66 ± 0.48, and 16.62 ± 0.56 dB, respectively. Our approach achieved higher SNR on this data set, and yielded 2.11 dB enhancement in SNR as compared to direct acquisition. Two-sample *t*-test at the 5% significance level showed that our approach significantly outperformed DA ($p = 3.02e^{-6}$) and SISR ($p = 2.40e^{-3}$). Wilcoxon signed-rank tests, where the null hypothesis was the difference of two sets of data comes from a distribution with zero median, showed that the population mean rank of our approach significantly differed from the two baselines in SNR at the 5% significance level (rejected the null hypothesis with $p = 7.8e^{-3}$ for both DA and SISR). **(B–D)** Our approach consistently offered the highest CNRs between the three types of brain tissues on this data set. In particular, our approach achieved 1.31 dB higher CNR between GM and WM than direct acquisition. The results show that SRR led to considerably improved SNR and CNR as compared to direct HR acquisition.

of data comes from a distribution with zero median, showed that the population mean rank of our approach significantly differed from the baselines in PVE at the 5% significance level (rejected the null hypothesis with $p = 8.86e^{-5}$ for IAA, $p = 2.76e^{-2}$ for TV, $p = 1.89e^{-4}$ for NLU, and $p = 2.93e^{-4}$ for SRCNN). **Figure 6B** shows the demonstration of the partial volume effect estimation on a representative image. The curve with a square marker shows the voxel distribution of the image. The dotted lines depict the three Gaussian components in the fitted GMM. The solid line addresses the fitted GMM. The three components from left to right represented the voxels from GM, WM, and CSF, respectively. The difference in the area under the curve between the voxel distribution and the fitted GMM in the range between two successive components corresponded to the estimate of the partial volume effect.

Figure 7 shows the estimated voxels suffering from partial volume averaging in the representative slice from the image reconstructed by the four baselines and our approach, respectively. The results show that almost all voxels with partial volume effect were from the boundaries between different types of brain tissues. Our approach comprised much fewer voxels with partial volume effect than the four baseline methods. The red arrows highlight the image regions with severe partial volume effect in the slice obtained from the four baseline methods. The results demonstrate that our approach offered considerably enhanced spatial resolution of this image.

Figure 8 shows the qualitative results in representative slices of the images reconstructed by the five SRR methods. The results show that our approach achieved the best qualitative performance with regarding to the image contrast and sharpness, in particular, on the delineation of the structures of the hippocampus as shown in the coronal and sagittal planes. The TV method sharpened the image excessively, resulting in noisy reconstructions. Our approach appropriately suppressed the noise contamination while enhancing the sharpness of the image edges. The images reconstructed by IAA, NLU, and SRCNN contained artifacts caused by averaging the images transformed due to the alignment, as highlighted by the red arrows. In contrast, our approach was not affected by the alignment.

4. DISCUSSION

We have developed a methodology to perform fast and high-resolution neonatal brain MRI. This methodology allows for high spatial resolution, high SNR and CNR, and reduced scan time, in comparison to direct HR acquisition. We have achieved high quality brain MRI at a spatial resolution of isotropic 0.4 mm with 6 min of imaging time. We have also demonstrated our approach on simulated data as well as clinical data acquired from twenty newborns. The experimental results have demonstrated that our

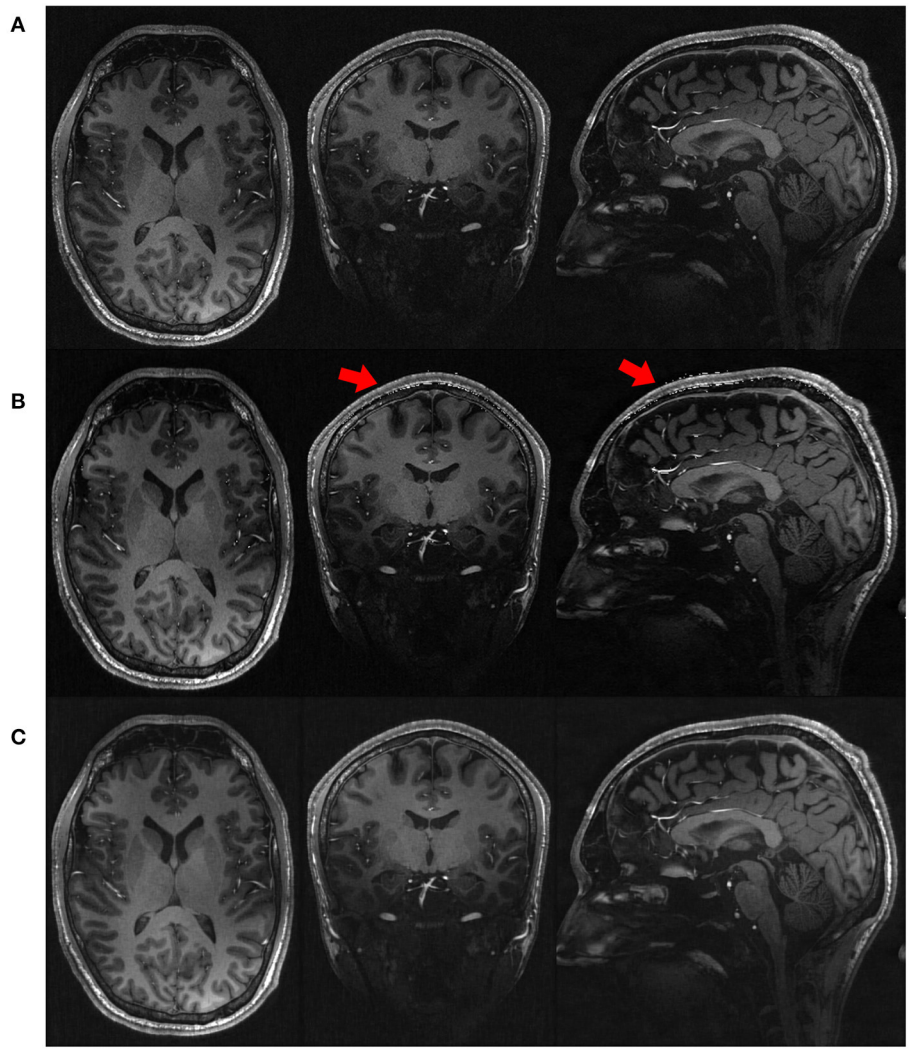


FIGURE 4 | Qualitative results in representative slices from the images **(A)** directly acquired, **(B)** reconstructed by SISR, and **(C)** reconstructed by our approach, respectively, on the simulated data set. The slices directly acquired and formed by SISR were much noisy as compared to our reconstructions. The noise was more prominent for SISR in the voxels from the skull, as highlighted by the red arrows. Although what we were interested in were CSF, GM, and WM, the noisy voxels from the skull rendered that SISR generated noise all over the images but just not as obvious as those from the skull.

approach allows for fast and high-quality neonatal brain MRI for both research and clinical studies.

We have shown in the simulation experiment that SRR achieved 7.7% lower partial volume effects and 2.11 dB higher SNR than the direct acquisitions at the resolution of isotropic 0.5 mm, while, in parallel, with only 68% of scan time of direct HR acquisition, as reported in **Figure 2**. Because the directly acquired HR images were very noisy, the image edges were blurred by the noise and in turn the spatial resolution was reduced. In SRR, because thick slices were used, the SNR was improved in the LR images, as described in the forward model shown in Equation (2). The blur kernel $\mathbf{H}_k$ in Equation (2) reduced the noise by the low-pass filtering. The thicker the slices, the more the reduction in the noise. Furthermore, the gradient guidance regularized SRR algorithm was used to reconstruct the HR images in both SISR and our approach. This algorithm incorporates an image deconvolutional filter that allows for further noise reduction in the HR reconstructions. If the scan time can be increased, e.g., taking the rest 32% of scan time to acquire more LR images with our protocol, our approach can achieve much higher spatial resolution and SNR.

In the experiment on the clinical T2 FSE data, we have shown that our approach generated comparable SNR to the IAA, NLU, and SRCNN methods while considerably higher spatial resolution. The averaging operation in the three baseline methods improved the SNR since the noise was smoothed out by the averaging. However, the averaging unexpectedly reduced the spatial resolution since it also blurred the tissue boundaries (image edges), as shown in **Figure 7**. Our approach, instead of averaging the data, combined the three LR images in a

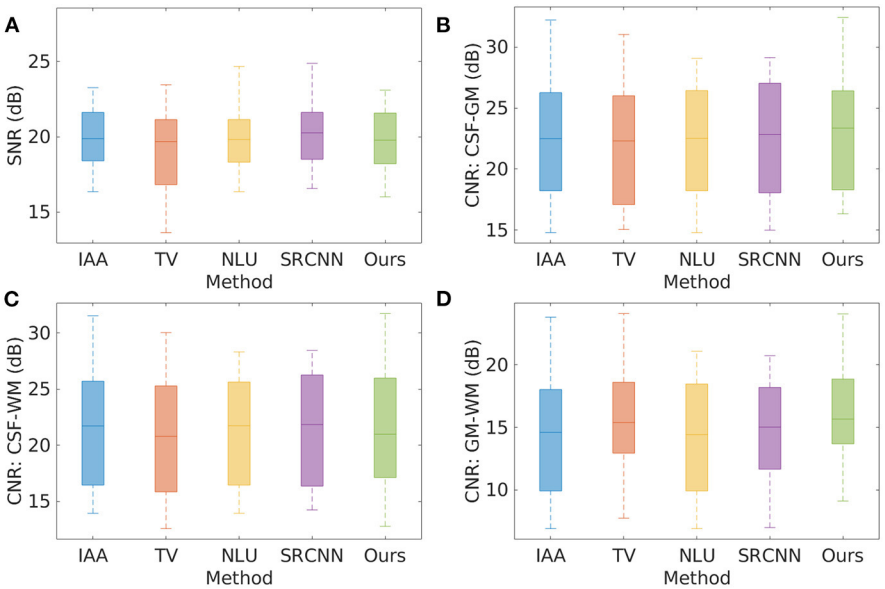


FIGURE 5 | Quality of the 20 HR images reconstructed by the five SRR methods on the clinical data set in terms of SNR and CNR. **(A)** SNR; **(B)** CNR between cerebrospinal fluid and gray matter; **(C)** CNR between cerebrospinal fluid and white matter; and **(D)** CNR between gray matter and white matter. The average SNR achieved by the five methods are, respectively: IAA = 20.19 ± 2.57 dB, TV = 19.17 ± 3.40 dB, NLU = 19.92 ± 2.04 dB, SRCNN = 20.18 ± 1.98 dB, Ours = 20.04 ± 2.77 dB. IAA, NLU, and SRCNN generated high SNR, as they benefited from the averaging to improve the SNR and CNR. Our approach offered comparable SNR with IAA, NLU, and SRCNN, and outperformed TV by ∼1 dB in terms of SNR. Our approach generated slightly superior CNRs to the five baselines about cerebrospinal fluid, and yielded considerably higher CNR between gray matter and white matter than these baselines.

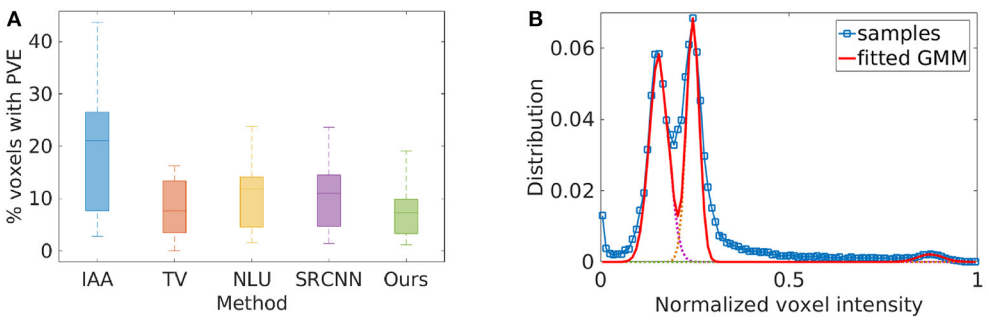


FIGURE 6 | Spatial resolution evaluated from the twenty images reconstructed by the five SRR methods on the clinical data set in terms of partial volume effect (PVE). **(A)** The average PVE achieved by the five methods are, respectively: IAA = 19.40 ± 11.85%, TV = 9.02 ± 7.30%, NLU = 11.35 ± 7.69%, SRCNN =10.88 ± 7.46%, Ours = 7.25 ± 4.37%. Our approach offered a considerably lower percentage of the voxels suffering from PVE in the HR reconstructions than the four baselines, leading to substantially enhanced spatial resolution. Two-sample *t*-test at the 5% significance level showed that our approach significantly outperformed IAA ($p = 1.40e^{-6}$), NLU ($p = 2.65e^{-4}$), and SRCNN ($p = 5.39e^{-4}$). Wilcoxon signed-rank tests, where the null hypothesis was the difference of two sets of data comes from a distribution with zero median, showed that the population mean rank of our approach significantly differed from the baselines in PVE at the 5% significance level (rejected the null hypothesis with $p = 8.86e^{-5}$ for IAA, $p = 2.76e^{-2}$ for TV, $p = 1.89e^{-4}$ for NLU, and $p = 2.93e^{-4}$ for SRCNN). **(B)** The demonstration of the PVE estimation on a representative image. The curve with a square marker shows the voxel distribution of the image. The dotted lines depict the three Gaussian components in the GMM. The solid line addresses the fitted GMM. The three components from left to right represented the voxels from GM, WM, and CSF, respectively. The difference in the area under the curve between the voxel distribution and the fitted GMM in the range between two successive components corresponded to the estimate of the PVE.

deconvolution manner that simultaneously improved the spatial resolution and SNR. As addressed in the forward model shown in Equation (2), the acquired image $\mathbf{y}_k$ was degraded by the convolution with the blur kernel $\mathbf{H}_k$. In the derived inverse problem defined in Equation (3), a deconvolution operation was leveraged, as an inverse operation of the convolution with $\mathbf{H}_k$, to restore the image from the blurring. This operation is also known as deblurring. As only the kernel $\mathbf{H}_k$ was involved, the noise that was filtered out in the convolution was not restored by the deconvolution, leading to an improved SNR in the reconstructed HR image. The TV method leveraged the same deconvolution scheme as our approach. However, the TV prior sharpened the

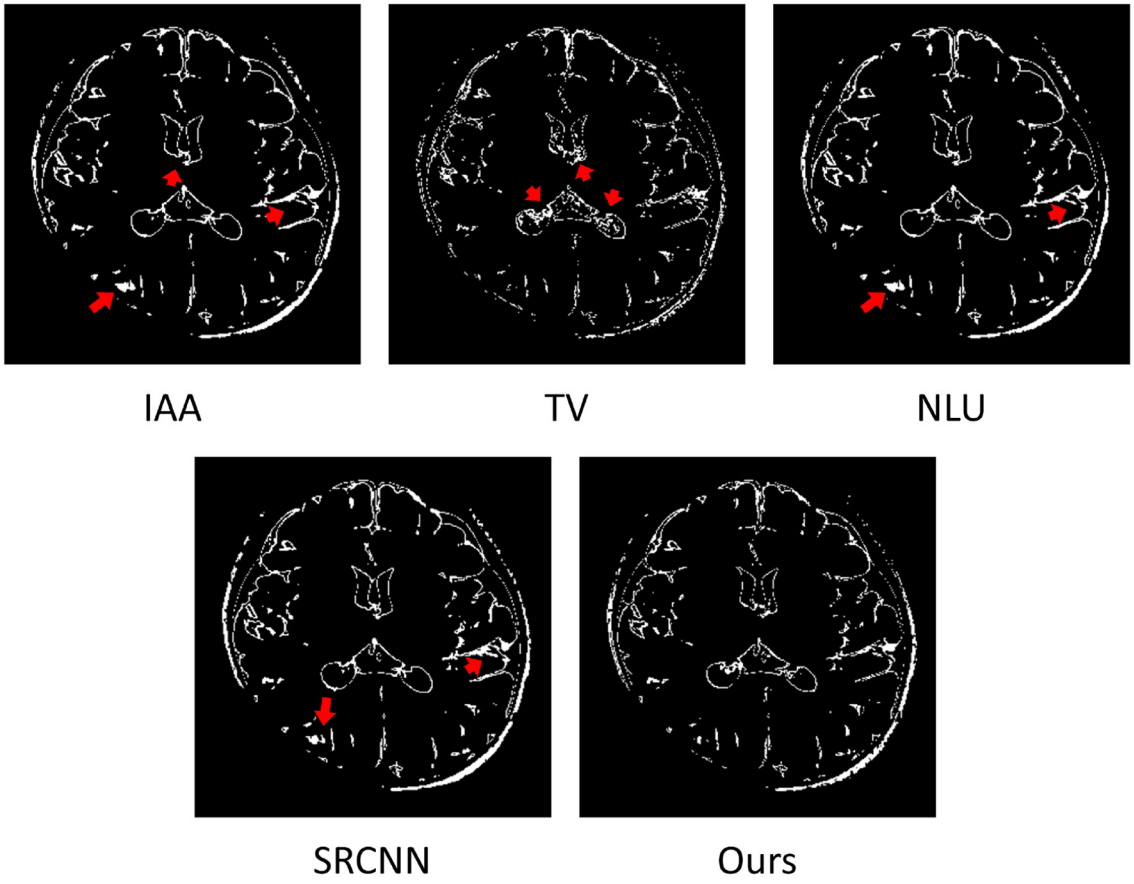


FIGURE 7 | Estimated voxels suffering from partial volume effect (PVE) in the representative slice of from the image reconstructed by the four baselines and our approach. The results show that almost all voxels with PVE were from the boundaries between different types of brain tissues. Our approach comprised much fewer voxels with PVE than the four baselines. The red arrows highlight the image regions with severe PVE in the slice obtained from the four baseline methods. The results demonstrate that our approach offered considerably enhanced spatial resolution of this image.

image excessively as it left the local smoothing unconsidered, resulting in noisy reconstructions.

As shown in **Figure 8**, the IAA, NLU, and SRCNN methods introduced the artifacts caused by averaging over the different number of voxels at the lattice because the image alignment rendered some voxels with undefined intensity values. Although the artifacts were outside the brain in this case, it would be an issue if the LR images, which only contain partial regions of the brain, are included in the SRR, leading to usable reconstructions. Benefiting from the method that we used to combine the images, our approach was not affected by the alignment and did not introduce such artifacts in the HR reconstructed images. As shown in the inverse problem defined in Equation (3) for our SRR, the deconvolution operations on each LR image are jointly combined in the data fidelity term. Furthermore, the regularization incorporates a spatial gradient guidance that constrains the HR reconstruction including local smooth regions separated by strong image edges. The ℓ_1-minimization imposed on the regularization guarantees that the local regions are not smoothed excessively. Consequently, our SRR offered both local region smoothing for homogeneous intensities and edge enhancement for tissue boundary preservation in the reconstructed HR image and did not involve the artifacts raised by the image alignment.

Our protocol allows for acquiring a T2 FSE image at the resolution of $0.39 \times 0.39 \times 2$ mm in 2 min. It is considerably fast for neonatal MRI to obtain an image with T_2 contrast at the resolution of isotropic 0.39 mm in 6 min of total imaging time. As a comparison, in 6 min of imaging time, we can only directly acquire a 3D T2 SPACE image at the resolution of isotropic 1 mm on our 3T scanner. Acquiring that same data at the resolution of isotropic 0.39 mm can be carried out, but acquires about 16.9 times more data, and so requires an extended acquisition time, with lengthened phase encodes, reduced readout bandwidth per pixel, and much more demanding variable flip angle calculation for signal loss in the lengthened phase encodes. Assuming we account for only the increased number of phase encodes required, this data would require $6 \times 6.57 = 39.4$ min to acquire. In addition, the SNR is reduced as each voxel shifts from 1 cubic mm to 0.39^3 cubic mm, a reduction in the signal by a factor of 16.9.

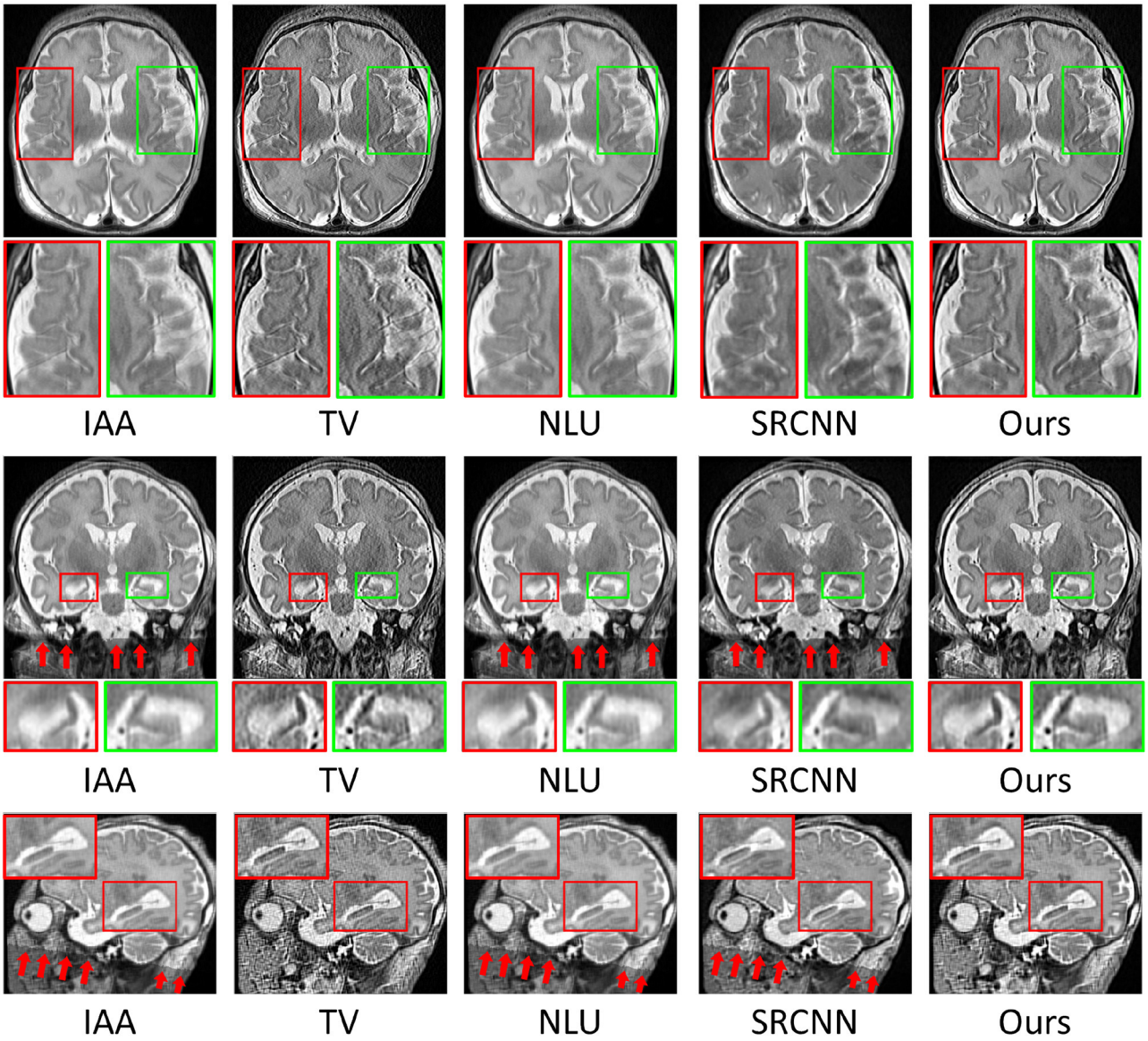


FIGURE 8 | Qualitative results in representative slices of the images reconstructed by the five SRR methods. Our approach achieved the best qualitative performance with regarding to the image contrast and sharpness, in particular, on the delineation of the structures of the hippocampus as shown in the coronal and sagittal planes. The TV method sharpened the image excessively, resulting in the noisy reconstruction. Our approach appropriately suppressed the noise contamination while enhancing the sharpness of the image edges. The images reconstructed by IAA, NLU, and SRCNN contained artifacts caused by averaging the images transformed due to the alignment, as highlighted by the red arrows. In contrast, our approach was not affected by the alignment.

In order for the HR data SNR to match the SNR of the 1 cubic mm data requires increasing the SNR by a factor of 16.9, which can be done by averaging together $16.9 \times 16.9 \approx 285$ acquisitions. Consequently acquired one HR image with matched SNR would require $285 \times 39.4 = 11{,}266$ min, or slightly shorter than 8 days in the MRI scanner.

Our approach enables extensive resolution critical clinical applications due to the enhanced spatial resolution and improved SNR while in parallel at reduced imaging time. It has shown that high spatial resolution facilitates the diagnosis of brain diseases, such as epilepsy (Conlon et al., 1988), multiple sclerosis (Truyen et al., 1996), and tumor characterization (Naruse et al., 1986). Our approach has achieved an isotropic spatial resolution of 0.4 mm, which allows for the clinical routines, such as the detection of signal abnormalities due to brain injury and the measurement of biometrics for impaired brain growth (Kidokoro et al., 2013), and in turn enables new assessment tools for neonatal brain MRI. Our fast and high-resolution imaging technique can be applied to the clinical and scientific research studies in the neonatal brain, such as the prediction and prognosis of brain injury (Kidokoro

et al., 2014; Haebich et al., 2019), for a better understanding of the potential pathways leading to altered brain structure and outcome in the preterm infant (Inder et al., 1999; Thompson et al., 2007).

In conclusion, we have exploited the acquisition strategy for improved SRR in neonatal brain MRI, which utilizes multiple anisotropic acquisitions with variable directions in slice selection. We have achieved neonatal brain MRI at a spatial resolution of isotropic 0.4 mm with 6 min of imaging time. We have demonstrated that our approach enabled considerably fast and high-quality neonatal brain MRI, as compared to direct HR acquisition. Extensive experimental results have shown that our approach allowed for high quality neonatal brain MRI for both scientific research and clinical studies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board, Boston Children's Hospital. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

YS: methodology, software, validation, formal analysis, and writing. SW: conceptualization, methodology, and supervision. AG: methodology and validation. OA: methodology, validation, and data curation. All authors contributed to the article and approved the submitted version.

FUNDING

Research reported in this publication was supported in part by the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke, the National Institute of Mental Health, and the Office Of The Director, National Institutes Of Health of the National Institutes of Health (NIH) under Award Numbers R01 NS079788, R01 EB019483, R01 EB018988, R01 NS106030, IDDRC U54 HD090255, S10OD025111; a research grant from the Boston Children's Hospital Translational Research Program; a Technological Innovations in Neuroscience Award from the McKnight Foundation; a research grant from the Thrasher Research Fund; and a pilot grant from National Multiple Sclerosis Society under Award Number PP-1905-34002. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, the Boston Children's Hospital, the Thrasher Research Fund, the National Multiple Sclerosis Society, or the McKnight Foundation.

REFERENCES

Afacan, O., Erem, B., Roby, D. P., Roth, N., Roth, A., Prabhu, S. P., et al. (2016). Evaluation of motion and its effect on brain magnetic resonance image quality in children. *Pediatr. Radiol.* 46, 1728–1735. doi: 10.1007/s00247-016-3677-9

Annink, K. V., van der Aa, N. E., Dudink, J., Alderliesten, T., Groenendaal, F., Lequin, M., et al. (2020). Introduction of ultra-high-field MR imaging in infants: preparations and feasibility. *Am. J. Neuroradiol.* 41, 1532–1537. doi: 10.3174/ajnr.A6702

Bertsekas, D. (1999). *Nonlinear Programming, 2nd Edn*. Athena Scientific.

Chaudhari, A., Fang, Z., Kogan, F., Wood, J., Stevens, K., Gibbons, E., et al. (2018). Super-resolution musculoskeletal MRI using deep learning. *Magn. Reson. Med.* 80, 2139–2154. doi: 10.1002/mrm.27178

Chen, Y., Shi, F., Christodoulou, A. G., Xie, Y., Zhou, Z., and Li, D. (2018). "Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Granada). doi: 10.1007/978-3-030-00928-1_11

Cherukuri, V., Guo, T., Schiff, S. J., and Monga, V. (2020). Deep MR brain image super-resolution using spatio-structural priors. *IEEE Trans. Image Process.* 29, 1368–1383. doi: 10.1109/TIP.2019.2942510

Conlon, P., Trimble, M., Rogers, D., and Callicott, C. (1988). Magnetic resonance imaging in epilepsy: a controlled study. *Epilepsy Res.* 2, 37–43. doi: 10.1016/0920-1211(88)90008-3

Dalca, A. V., Bouman, K. L., Freeman, W. T., Rost, N. S., Sabuncu, M. R., and Golland, P. (2019). Medical image imputation from image collections. *IEEE Trans. Med. Imaging* 38, 504–514. doi: 10.1109/TMI.2018.2866692

Daubechies, I., Defrise, M., and Mol, C. D. (2003). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.* 57, 1413–1457. doi: 10.1002/cpa.20042

Ding, Y., Acosta, R., Enguix, V., Suffren, S., Ortmann, J., Luck, D., et al. (2020). Using deep convolutional neural networks for neonatal brain image segmentation. *Front. Neurosci.* 14:207. doi: 10.3389/fnins.2020.00207

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 295–307. doi: 10.1109/TPAMI.2015.2439281

Dubois, J., Alison, M., Counsell, S., Lucie, H.-P., Hüppi, P. S., and Benders, M. (2020). MRI of the neonatal brain: a review of methodological challenges and neuroscientific advances. *J. Magn. Reson. Imaging* 53, 1318–1343. doi: 10.1002/jmri.27192

Dubois, J., Lefévre, J., Angleys, H., Leroy, F., Fischer, C., Lebenberg, J., et al. (2019). The dynamics of cortical folding waves and prematurity-related deviations revealed by spatial and spectral analysis of gyrification. *NeuroImage* 185, 934–946. doi: 10.1016/j.neuroimage.2018.03.005

Fiat, D. (2001). *Method of Enhancing an MRI Signal*. US Patent 6,294,914.

Gholipour, A., Afacan, O., Aganj, I., Scherrer, B., Prabhu, S. P., Sahin, M., and Warfield, S. K. (2015). Super-resolution reconstruction in frequency, image, and wavelet domains to reduce through-plane partial voluming in MRI. *Med. Phys.* 42, 6919–6932. doi: 10.1118/1.4935149

Gholipour, A., Estroff, J. A., Sahin, M., Prabhu, S. P., and Warfield, S. K. (2010a). "Maximum a posteriori estimation of isotropic high-resolution volumetric MRI from orthogonal thick-slice scans," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Berlin; Heidelberg: Springer), 109–116. doi: 10.1007/978-3-642-15745-5_14

Gholipour, A., Estroff, J. A., and Warfield, S. K. (2010b). Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain MRI. *IEEE Trans. Med. Imaging* 29, 1739–1758. doi: 10.1109/TMI.2010.2051680

Giampietri, M., Bartalena, L., Guzzetta, A., Boldrini, A., and Ghirri, P. (2015). New techniques in the study of the brain development in newborn. *Front. Hum. Neurosci.* 8:1069. doi: 10.3389/fnhum.2014.01069

Greenspan, H., Oz, G., Kiryati, N., and Peled, S. (2002). MRI inter-slice reconstruction using super-resolution. *Magn. Reson. Imaging* 20, 437–446. doi: 10.1016/S0730-725X(02)00511-8

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., et al. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magn. Reson. Med.* 47, 1202–1210. doi: 10.1002/mrm.10171

Haebich, K. M., Willmott, C., Scratch, S., Pascoe, L., Lee, K., Spencer-Smith, M. M., et al. (2019). Neonatal brain abnormalities and brain volumes associated with goal setting outcomes in very preterm 13-year-olds. *Brain Imaging Behav.* 14, 1062–1073. doi: 10.1007/s11682-019-00039-1

Hansen, M. S., and Kellman, P. (2015). Image reconstruction: an overview for clinicians. *J. Magn. Reson. Imaging* 41, 573–585. doi: 10.1002/jmri.24687

Inder, T., Huppi, P., Warfield, S., Kikinis, R., Zientara, G., Barnes, P., et al. (1999). Periventricular white matter injury in the premature infant is followed by reduced cerebral cortical gray matter volume at term. *Ann. Neurol.* 46, 755–760. doi: 10.1002/1531-8249(199911)46:5<755::AID-ANA11>3.0.CO;2-0

Jiang, S., Xue, H., Glover, A., Rutherford, M., Rueckert, D., and Hajnal, J. (2007). MRI of moving subjects using multislice snapshot images with volume reconstruction (SVR): application to fetal, neonatal, and adult brain studies. *IEEE Trans. Med. Imaging* 26, 967–980. doi: 10.1109/TMI.2007.895456

Kainz, B., Steinberger, M., Wein, W., Murgasova, M., Malamateniou, C., Keraudren, K., et al. (2015). Fast volume reconstruction from motion corrupted stacks of 2d slices. *IEEE Trans. Med. Imaging* 34, 1901–1913. doi: 10.1109/TMI.2015.2415453

Kidokoro, H., Anderson, P., Doyle, L., Woodward, L., Neil, J., and Inder, T. (2014). Brain injury and altered brain growth in preterm infants: predictors and prognosis. *Pediatrics* 134, e444–e453. doi: 10.1542/peds.2013-2336

Kidokoro, H., Neil, J., and Inder, T. (2013). New MR imaging assessment tool to define brain abnormalities in very preterm infants at term. *Am. J. Neuroradiol.* 34, 2208–2214. doi: 10.3174/ajnr.A3521

Laidlaw, D., Fleischer, K., and Barr, A. H. (1998). Partial-volume Bayesian classification of material mixtures in MR volume data using voxel histograms. *IEEE Trans. Med. Imaging* 17, 74–86. doi: 10.1109/42.668696

Langford, E. (2006). Quartiles in elementary statistics. *J. Stat. Educ.* 14, 1–18. doi: 10.1080/10691898.2006.11910589

Lusebrink, F., Sciarra, A., Mattern, H., Yakupov, R., and Speck, O. (2017). T1-weighted *in vivo* human whole brain MRI dataset with an ultrahigh isotropic resolution of 250 um. *Sci. Data* 4:170032. doi: 10.1038/sdata.2017.62

Makropoulos, A., Robinson, E., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., et al. (2018). The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage* 173, 88–112. doi: 10.1016/j.neuroimage.2018.01.054

Manjón, J. V., Coupé, P., Buades, A., Fonov, V. S., Collins, D. L., and Robles, M. (2010). Non-local MRI upsampling. *Med. Image Anal.* 14, 784–792. doi: 10.1016/j.media.2010.05.010

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of boxplot. *Am. Stat.* 32, 12–16. doi: 10.1080/00031305.1978.10479236

Mongerson, C. R. L., Wilcox, S. L., Goins, S. M., Pier, D., Zurakowski, D., Jennings, R. W., et al. (2019). Infant brain structural MRI analysis in the context of thoracic non-cardiac surgery and critical care. *Front. Pediatr.* 7:315. doi: 10.3389/fped.2019.00315

Murgasova, M., Quaghebeur, G., Rutherford, M., Hajnal, J., and Schnabel, J. A. (2012). Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Med. Image Anal.* 16, 1550–1564. doi: 10.1016/j.media.2012.07.004

Naruse, S., Horikawa, Y., Tanaka, C., Hirakawa, K., Nishikawa, H., and Yoshizaki, K. (1986). Significance of proton relaxation time measurement in brain edema, cerebral infarction and brain tumors. *Magn. Reson. Imaging* 4, 293–304. doi: 10.1016/0730-725X(86)91039-8

Peled, S., and Yeshurun, Y. (2002). Superresolution in MRI-perhaps sometimes. *Magn. Reson. Med.* 48:409. doi: 10.1002/mrm.10237

Plenge, E., Poot, D. H. J., Bernsen, M., Kotek, G., Houston, G., Wielopolski, P., et al. (2012). Super-resolution methods in MRI: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magn. Reson. Med.* 68, 1983–1993. doi: 10.1002/mrm.24187

Poot, D., Van Meir, V., and Sijbers, J. (2010). "General and efficient super-resolution method for multi-slice MRI," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Beijing). doi: 10.1007/978-3-642-15705-9_75

Pruessmann, K. P., Weiger, M., Scheidegger, M. B., and Boesiger, P. (1999). SENSE: sensitivity encoding for fast MRI. *Magn. Reson. Med.* 42, 952–962. doi: 10.1002/(SICI)1522-2594(199911)42:5<952::AID-MRM16>3.0.CO;2-S

Rousseau, F., Glenn, O., Iordanova, B., Rodríguez-Carranza, C., Vigneron, D., Barkovich, A., et al. (2005). A novel approach to high resolution fetal brain MR imaging. *Med. Image Comput. Comput. Assist. Interv.* 8(Pt 1), 548–555. doi: 10.1007/11566465_68

Rousseau, F., Kim, K., Studholme, C., Koob, M., and Dietemann, J. (2010). On super-resolution for fetal brain MRI. *Med. Image Comput. Comput. Assist. Interv.* 13(Pt 2), 355–362. doi: 10.1007/978-3-642-15745-5_44

Scheffler, K. (2002). Superresolution in MRI? *Magn. Reson. Med.* 48:408. doi: 10.1002/mrm.10203

Scherrer, B., Afacan, O., Taquet, M., Prabhu, S. P., Gholipour, A., and Warfield, S. K. (2015). "Accelerated high spatial resolution diffusion-weighted imaging," in *International Conference on Information Processing in Medical Imaging (IPMI)* (Skye). doi: 10.1007/978-3-319-19992-4_6

Scherrer, B., Gholipour, A., and Warfield, S. K. (2012). Super-resolution reconstruction to increase the spatial resolution of diffusion weighted images from orthogonal anisotropic acquisitions. *Med. Image Anal.* 16, 1465–1476. doi: 10.1016/j.media.2012.05.003

Shi, F., Cheng, J., Wang, L., Yap, P., and Shen, D. (2015). LRTV: MR image super-resolution with low-rank and total variation regularizations. *IEEE Trans. Med. Imaging* 34, 2459–2466. doi: 10.1109/TMI.2015.2437894

Shilling, R. Z., Robbie, T. Q., Bailloeul, T., Mewes, K., Mersereau, R. M., and Brummer, M. E. (2009). A super-resolution framework for 3-D high-resolution and high-contrast imaging using 2-D multislice MRI. *IEEE Trans. Med. Imaging* 28, 633–644. doi: 10.1109/TMI.2008.2007348

Sui, Y., Afacan, O., Gholipour, A., and Warfield, S. K. (2019). "Isotropic MRI super-resolution reconstruction with multi-scale gradient field prior," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Shenzhen). doi: 10.1007/978-3-030-32248-9_1

Sui, Y., Afacan, O., Gholipour, A., and Warfield, S. K. (2020). "Learning a gradient guidance for spatially isotropic MRI super-resolution reconstruction," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI), Vol. 11766* (Lima), 3–11. doi: 10.1007/978-3-030-59713-9_14

Thompson, D., Warfield, S., Carlin, J., Pavlović, M., Wang, H., Bear, M. J., et al. (2007). Perinatal risk factors altering regional brain structure in the preterm infant. *Brain* 130(Pt 3), 667–677. doi: 10.1093/brain/awl277

Tortora, D., Severino, M., Biase, C. D., Malova, M., Parodi, A., Minghetti, D., et al. (2019). Early pain exposure influences functional brain connectivity in very preterm neonates. *Front. Neurosci.* 13:899. doi: 10.3389/fnins.2019.00899

Tourbier, S., Bresson, X., Hagmann, P., Thiran, J., Meuli, R., and Cuadra, M. (2015). An efficient total variation algorithm for super-resolution in fetal brain MRI with adaptive regularization. *Neuroimage* 118, 584–597. doi: 10.1016/j.neuroimage.2015.06.018

Truyen, L., van Waesberghe, J. H. V., van Walderveen, M. V., van Oosten, B., Polman, C., Hommes, O., et al. (1996). Accumulation of hypointense lesions ("black holes") on t1 spin-echo MRI correlates with disease progression in multiple sclerosis. *Neurology* 47, 1469–1476. doi: 10.1212/WNL.47.6.1469

Tsai, R., and Huang, T. (1984). "Multi-frame image restoration and registration," in *Advances in Computer Vision and Image Processing*.

Van Reeth, E., Tham, I. W., Tan, C. H., and Poh, C. L. (2012). Super-resolution in magnetic resonance imaging: a review. *Concepts Magn. Reson.* 40, 306–325. doi: 10.1002/cmr.a.21249

Wang, J., Chen, Y., Wu, Y., Shi, J., and Gee, J. (2020). "Enhanced generative adversarial network for 3d brain MRI super-resolution," in *IEEE Winter Conference on Applications of Computer Vision* (Aspen, CO), 3627–3636. doi: 10.1109/WACV45572.2020.9093603

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error measurement to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Weisenfeld, N., and Warfield, S. (2009). Automatic segmentation of newborn brain MRI. *NeuroImage* 47, 564–572. doi: 10.1016/j.neuroimage.2009.04.068

Xue, X., Wang, Y., Li, J., Jiao, Z., Ren, Z., and Gao, X. (2020). Progressive sub-band residual-learning network for MR image super-resolution. *IEEE J. Biomed. Health Inform.* 24, 377–386. doi: 10.1109/JBHI.2019.2945373

Zhao, X., Zhang, Y., Zhang, T., and Zou, X. (2019). Channel splitting network for single MR image super-resolution. *IEEE Trans. Image Process.* 28, 5649–5662. doi: 10.1109/TIP.2019.2921882

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fnins.2021.674719

The Influence of Radio-Frequency Transmit Field Inhomogeneities on the Accuracy of G-ratio Weighted Imaging

*Tim M. Emmenegger[1,2], Gergely David[1,2], Mohammad Ashtarayeh[2], Francisco J. Fritz[2], Isabel Ellerbrock[2,3], Gunther Helms[4], Evelyne Balteau[5], Patrick Freund[1,6,7] and Siawoosh Mohammadi[2,6]**

[1] Spinal Cord Injury Center Balgrist, University Hospital Zurich, University of Zurich, Zurich, Switzerland, [2] Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [3] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden, [4] Medical Radiation Physics, Clinical Sciences Lund (IKVL), Lund University, Lund, Sweden, [5] GIGA Institute, University of Liège, Liège, Belgium, [6] Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, [7] Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

OPEN ACCESS

Edited by:
Tim B. Dyrby, Technical University of Denmark, Denmark

Reviewed by:
Viktor Vegh, The University of Queensland, Australia
Olivier Commowick, Inria Rennes–Bretagne Atlantique Research Centre, France

***Correspondence:**
Siawoosh Mohammadi
s.mohammadi@uke.de

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *01 March 2021*
Accepted: *01 June 2021*
Published: *05 July 2021*

Citation:
Emmenegger TM, David G, Ashtarayeh M, Fritz FJ, Ellerbrock I, Helms G, Balteau E, Freund P and Mohammadi S (2021) The Influence of Radio-Frequency Transmit Field Inhomogeneities on the Accuracy of G-ratio Weighted Imaging. Front. Neurosci. 15:674719. doi: 10.3389/fnins.2021.674719

G-ratio weighted imaging is a non-invasive, *in-vivo* MRI-based technique that aims at estimating an aggregated measure of relative myelination of axons across the entire brain white matter. The MR g-ratio and its constituents (axonal and myelin volume fraction) are more specific to the tissue microstructure than conventional MRI metrics targeting either the myelin or axonal compartment. To calculate the MR g-ratio, an MRI-based myelin-mapping technique is combined with an axon-sensitive MR technique (such as diffusion MRI). Correction for radio-frequency transmit (B1+) field inhomogeneities is crucial for myelin mapping techniques such as magnetization transfer saturation. Here we assessed the effect of B1+ correction on g-ratio weighted imaging. To this end, the B1+ field was measured and the B1+ corrected MR g-ratio was used as the reference in a Bland-Altman analysis. We found a substantial bias (≈-89%) and error (≈37%) relative to the dynamic range of g-ratio values in the white matter if the B1+ correction was not applied. Moreover, we tested the efficiency of a data-driven B1+ correction approach that was applied retrospectively without additional reference measurements. We found that it reduced the bias and error in the MR g-ratio by a factor of three. The data-driven correction is readily available in the open-source hMRI toolbox (www.hmri.info) which is embedded in the statistical parameter mapping (SPM) framework.

Keywords: myelin volume fraction, axon volume fraction, radio-frequency transmit field inhomogeneities, B_1+ correction, multi-parameter mapping, diffusion MRI, magnetization transfer saturation, MR g-ratio

INTRODUCTION

The g-ratio [i.e., the ratio between the inner (r) and outer (R) radius of an axon with myelin sheath (g-ratio = r/R)] of a given axon quantifies the degree of relative myelination, ranging between 0 (no axon) and 1 (no myelin). The g-ratio captures both axonal and myelin damage by incorporating axonal and myelin volumes in one metric, making it potentially more specific to tissue integrity than

focusing on one of these aspects only. For example, in multiple sclerosis, the g-ratio increases if the underlying disease mechanism is solely driven by demyelination (Yu et al., 2019), but is expected to remain unaffected if demyelination is accompanied by axonal degeneration. To differentiate such processes and understand their functional implications, neuroscience and clinical research would greatly benefit from *in-vivo* whole-brain measurements of MR g-ratio. Until recently, the g-ratio was measurable only by means of histology (Hildebrand and Hahn, 1978), which restricted the analyses to a small number of axons and a limited number of small brain regions or pathways. Stikov et al. (2011, 2015) introduced a methodology for an MRI-based whole-brain "aggregate" g-ratio mapping, to which we refer as "MR g-ratio" or "g-ratio weighted imaging." In g-ratio weighted imaging, the MR g-ratio is computed on a voxel-by-voxel basis from the axonal (AVF) and myelin volume fraction (MVF) maps and reflects a weighted mean of g-ratio values within the voxel (West et al., 2016). Therefore, g-ratio weighted imaging requires the acquisition of separate sets of images that are sensitive to AVF and MVF, respectively (Campbell et al., 2018; Mohammadi and Callaghan, 2020). To generate MVF and AVF from the measured MR parameters, a calibration step is required that converts the measured MR-visible water signals into the respective volume fractions (Mohammadi and Callaghan, 2020).

Magnetization transfer saturation (MT_{sat}) has often been used as proxy for MVF (Mohammadi et al., 2015) as it is minimally affected by the longitudinal relaxation time (Helms et al., 2008) and is expected to show high correlation with macromolecular content (Sereno et al., 2013; Callaghan et al., 2015a; Campbell et al., 2018), making it a sensitive metric of MVF. One common approach to estimate AVF complements the parameters from neurite orientation and dispersion density imaging (NODDI Zhang et al., 2012) with a MVF-proxy, e.g., MT_{sat} (Ellerbrock and Mohammadi, 2018; Kamagata et al., 2019), to correct for the missing myelin water signal in diffusion MRI measurements (Stikov et al., 2015). Maps of MT_{sat} can be obtained, among others, from the multi-parameter mapping (MPM) protocol (Weiskopf et al., 2013) in combination with the hMRI toolbox[1] (Callaghan et al., 2019; Tabelow et al., 2019).

Although the MT_{sat} measure is largely insensitive to transmit field (B_1+) inhomogeneities (Helms et al., 2008), it still shows a residual dependence which introduces a bias and/or error in the MT_{sat} maps that can propagate into the MR g-ratio and lead to systematic bias. Such B_1+ inhomogeneities can be corrected based on an independently acquired B_1+ field map measurement (Helms, 2015; Helms et al., 2021). Residual B_1+ inhomogeneity effects on MT_{sat} have been shown to be not negligible when the B_1+ correction was omitted (Helms, 2015; Helms et al., 2021). However, the impact of B_1+ correction on MR g-ratio estimates is unknown. Additionally, it is unclear whether these residual B_1+ inhomogeneity in MT_{sat} and the MR g-ratio can retrospectively be corrected using a data-driven B_1+ field inhomogeneities estimation approach such as the "unified segmentation based correction of R1 maps for B_1+ inhomogeneities" (UNICORT, (Weiskopf et al., 2011)).

[1]www.hMRI.info

In this study, we investigate the effect of B_1+ inhomogeneities on MR g-ratio maps when omitting the B_1+ correction. As a reference, we use the B_1+ corrected MR g-ratio from a dataset of healthy controls. We compare the reference MR g-ratio values against (i) values obtained without B_1+ correction and (ii) values obtained with B_1+ correction using the data-driven UNICORT approach.

MATERIALS AND METHODS

Subjects

This study included 25 healthy control subjects (12 females, age (mean ± standard deviation) of 25.4 ± 2.4 years). They were recruited at the University Medical Centre Hamburg-Eppendorf and screened for neurological or psychiatric illness. The study was in agreement with the Declaration of Helsinki and was approved by the local ethics committee (Ärztekammer Hamburg #PV5141).

Data Acquisition

Each subject was scanned twice within 1 week in a whole-body 3T Tim TRIO MR scanner (Siemens Healthcare, Erlangen, Germany) using the body RF-coil for transmission and a 32-channel radiofrequency (RF) head coil for signal reception, respectively. The MR acquisition on both scan days included a multi-parameter mapping (MPM) (Weiskopf et al., 2013; Callaghan et al., 2015b) and a diffusion-weighted imaging (DWI) protocol. The MPM protocol consists of three differently weighted 3D-multi-echo spoiled gradient echo sequences (Siemens FLASH). The echo train length and flip angle for the proton density (PD) weighted, T1-weighted, and magnetization transfer (MT) weighted sequences were 8/6, 8/21, and 6/6°, respectively. The MT-weighted sequence had a Gaussian RF pulse (2 kHz off resonance with 4 ms duration and a nominal flip angle of 220°). All other sequence parameters were the same for the three sequences: repetition time (TR) 25 ms, echo spacing, resolution 0.8 mm isotropic; field of view (FoV) 166 × 224 × 256 mm^3, readout bandwidth 488 Hz/pixel, partially parallel imaging using the GRAPPA algorithm was employed in each phase-encoded direction (anterior-posterior and right-left) with 40 reference lines and a speed up factor of two, total acquisition time: ∼25 min. The B_1+ field reference map was acquired using the three-dimensional echo-planar imaging (3D EPI) method, including field maps for distortion correction (Lutti et al., 2010).

The DWI sequence was a twice-refocused single-shot spin-echo EPI scheme (Reese et al., 2003), consisting of 12 non-diffusion-weighted images (b_0 images), equidistantly distributed across the diffusion weighted images. The diffusion-weighted images were acquired at two *b*-values ($1000\frac{s}{mm^2}$ and $2000\frac{s}{mm^2}$), sampled along 60 unique diffusion-gradient directions within each shell. The entire protocol was repeated with identical parameters but with reversed phase encoding direction (anterior-posterior) to correct for susceptibility-related image distortions (blip-up, blip-down correction). In total, 264 images were acquired per subject (120 diffusion-weighted images, 12 b_0 images, each acquired twice). Other acquisition parameters were:

86 slices with no gap, TR = 7.1 s, TE = 122 ms, an isotropic voxel size of (1.6 mm)^3, FoV = 224 × 224 × 138 mm^3, 7/8 partial Fourier imaging in phase encoding direction, readout bandwidth. To accelerate the data acquisition, GRAPPA (in-plane acceleration with factor two) and simultaneous multi-slice acquisitions ("multiband," slice acceleration factor two) (Feinberg et al., 2010; Moeller et al., 2010; Xu et al., 2013) were used as described in Setsompop et al. (2012). The image reconstruction algorithm was provided by the University of Minnesota Centre for Magnetic Resonance Research. The total acquisition time was ∼37 min.

Data Processing

MT_{sat} maps were generated in the SPM-based hMRI toolbox (Tabelow et al., 2019). Note that the hMRI toolbox also generates additional maps of longitudinal (R_1) and effective transverse relaxation rates (R_2^*) and PD. Three MT_{sat} maps were generated: (i) MT_{sat}^{NO} maps, without B_1+ correction; (ii) MT_{sat}^{B1} map, using the reference B_1+ field map for correction (Lutti et al., 2010); and (iii) MT_{sat}^{UN} maps, using the data-driven UNICORT approach for B_1+ estimation (Weiskopf et al., 2011; see **Supplementary Figure 2**). UNICORT is a probabilistic framework for unified-segmentation based correction of R_1 maps for B_1+ inhomogeneities. The framework incorporates a physically informed generative model of smooth B_1+ inhomogeneities and their multiplicative effect on R_1 estimates (Weiskopf et al., 2011). Parameters used in UNICORT such as the smoothness and regularization were optimized for R_1 B_1+ correction in a 3T scanner (i.e., Tim Trio scanner—Weiskopf et al., 2011).

For B_1+ correction, we used the following heuristic correction factor as detailed in Helms (2015), and Helms et al. (2021):

$$MT_{sat}^{Corr} = MT_{sat}^{NO} \frac{1 - C}{1 - CB_1^+}, \quad (1)$$

where C has been calibrated to be 0.4 for the MT pulse used in this paper. B_1+ can be either measured ($MT_{sat}^{Corr} = MT_{sat}^{B1}$) or estimated with the UNICORT approach ($MT_{sat}^{Corr} = MT_{sat}^{UN}$).

The DWI data were processed based on the pipeline described in Ellerbrock and Mohammadi (2018) using the SPM-based ACID toolbox[2]. It included several artifact corrections such as Rician signal bias correction (i.e., denoising) (André et al., 2014), correction for eddy current and motion artifacts (Mohammadi et al., 2010, 2014), and correction for image distortions due to susceptibility artifact using reversed phase encoding (Ruthotto et al., 2012, 2013; Macdonald and Ruthotto, 2018). The corrected images were fitted with the NODDI signal model (Zhang et al., 2012) to estimate the intra-cellular volume fraction (ν_{icvf}), the isotropic volume fraction (ν_{iso}), and the orientation dispersion index (ODI) in each voxel.

[2]http://www.diffusiontools.com

Spatial Alignment

Co-registration

The voxel-wise arithmetic between the MT_{sat} and ν_{icvf} maps, necessary for MR g-ratio computation, requires an accurate spatial alignment between the two maps (Mohammadi et al., 2015). To this end, we created two white matter (WM) tissue probability maps (TPMs) based on the ODI and MT_{sat}^{B1} maps, respectively (**Figure 1**). To reduce the influence of contrast-specific artifacts (e.g., due to subject motion) on the registration quality, the WM TPM of the ODI map was co-registered to the WM TPM of the MT_{sat}^{B1} map using rigid-body registration (*spm_coreg* algorithm, SPM toolbox). The estimated transformation parameters were applied to all other NODDI maps as well. Note that the segmentation quality of the second session was unsatisfactory for two subjects, and the $R_1^{B_1}$ map (R_1 with B_1+ inhomogeneities bias correction using the B_1+ reference measurements) was used to generate the WM TPM instead. In another subject, the ν_{iso} was segmented instead of the ODI to achieve satisfactory WM segments.

Normalization

Spatial normalization was performed in four steps. First, a rough alignment of the MT_{sat}^{B1} maps with the T1-weighted MNI template image was achieved using the Auto-Reorient function (hMRI toolbox) and this was applied on the NODDI maps as well. Second, both MT_{sat}^{B1} maps of each subject (corresponding to two sessions) were registered to the mid-point average using the Pairwise Longitudinal Registration (SPM12). Hereby, values below zero and above 10 were excluded to improve the registration. Third, the resulting mid-point average image was normalized to the MNI space using the DARTEL-based (Ashburner, 2007) Spatial Processing module (hMRI toolbox). Fourth, a combined deformation field was generated per subject and session, combining the deformation fields from steps 2 and 3.

Computation of MVF_{MR}, AVF_{MR} and g_{MR}

In this section, our approach to estimating MVF and AVF from the measured MR parameters is introduced. The MR-based MVF (MVF_{MR}) was assumed to be proportional to MT_{sat} without intercept, following (Mohammadi and Callaghan, 2020):

$$MVF_{MR} = \alpha MT_{sat} \quad (2)$$

The proportionality constant α was estimated from Equation (2) in a region where the histological MVF (MVF_{hist}) was known. Due to the lack of own histological data, we used published histological data which contain the frequency distribution of inner-axon radius (r) and myelin sheath thickness (m) of 2,400 myelinated fibers in the medullary pyramids of a 71 years old human (see **Table 1** in Graf von Keyserlingk and Schramm, 1984). The total volume (TV) of the sample is the sum of the total volume of myelinated axons (TAV_m), unmyelinated axons (TAV_u), myelin volume (TMV), and extra-cellular volume (TEV). TAV_m was calculated as $\sum_{i=1}^{N_m} \pi r_i^2$ with i indexing the N_m myelinated axons only, and TMV was computed as $\sum_{i=1}^{N_m} \pi(r_i + m_i)^2 - TAV_m$. TAV_u, while not reported in Graf von Keyserlingk and Schramm (1984), was found to be

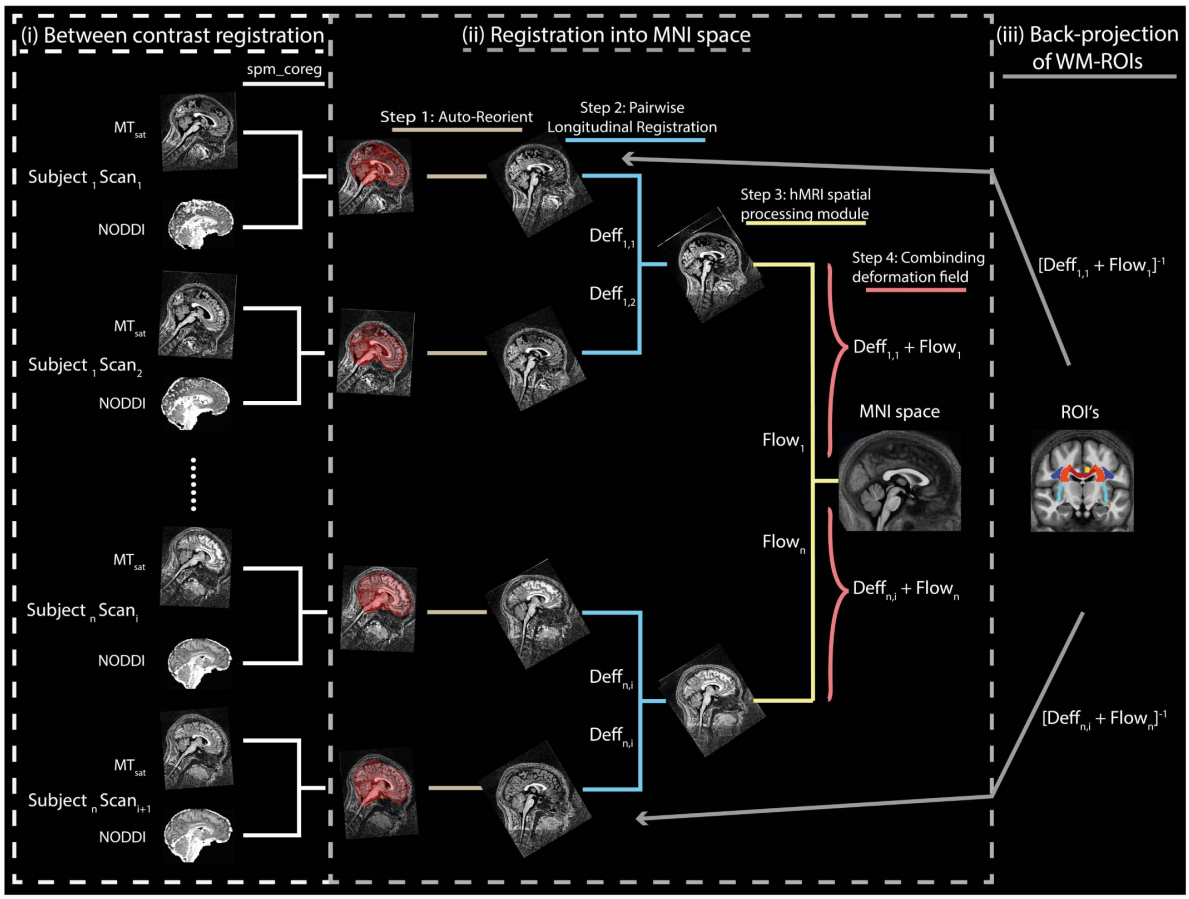


FIGURE 1 | Illustration of the spatial alignment pipeline of the MT_{sat} and NODDI maps. The pipeline consists of (i) co-registration between MT_{sat} and NODDI maps (driven by ODI map), (ii) normalization into MNI space, and (iii) back-projection of ROIs into the native space. Note that each subject consists of two sets of images acquired in separate sessions. In the co-registration step (section "Co-registration"), the white matter (WM) tissue probability map (TPM) of the ODI was co-registered to the WM TMP of the MT_{sat} in each subject and session using rigid-body registration (*spm_coreg* algorithm, SPM12). The resulting transformation was applied to all other NODDI maps as well. In the normalization step (section "Normalization"), MT_{sat} maps were roughly aligned with the T1-weighted MNI template in each subject and session using the *Auto-Reorient* function. The realigned MT_{sat} maps from both sessions were then registered to their mid-point average using the Pairwise Longitudinal Registration (SPM12). In each subject, the mid-point average MT_{sat} map was normalized to the MNI space using the DARTEL-based (Ashburner, 2007) Spatial Processing module. Finally, all deformation fields were converted to a single deformation field and applied on the NODDI maps. In the last step (section "Region of Interest Selection"), the ROIs and the WM masks were back-projected into the native space using the inverse of the combined deformation field.

approximately 43% of TAV_m for multiple mammals (Swadlow et al., 1980; LaMantia and Rakic, 1990; Olivares et al., 2001; Wang et al., 2008; Liewald et al., 2014). Note that the aforementioned papers typically reported the unmyelinated axons as 30% of the total volume of axons, which corresponds to 43% ($= \frac{0.3}{1-0.3} \cdot 100$) of TAV_m. EVF was estimated to be 25%, according to Lehmenkühler et al. (1993), Nicholson and Hrabìtová (2017), Tønnesen et al. (2018). Finally, MVF was calculated as

$$\mathrm{MVF_{hist}} \approx \frac{1}{\mathrm{TV}} \sum_{j=1}^{\mathrm{N}} \pi \left(\left(\mathrm{r}_j + \mathrm{m}_j \right)^2 - \mathrm{r}_j^2 \right) \tag{3}$$

with j indexing all N fibers, yielding $MVF_{hist} \approx 0.3623$. Plugging this value into Equation (2) (assuming that $MVF_{MR} \approx MVF_{hist}$) along with the group-average MT_{sat} within the medullary pyramids (see **Figure 2** for ROI definition) yielded an α of 0.2496 for MT_{sat}^{B1}, 0.2414 for MT_{sat}^{UN}, and 0.2884 for MT_{sat}^{NO}.

The MR-based AVF ($AVF_{MR} = (1-MVF_{MR})\ AWF_{MR}$) was calculated as

$$\mathrm{AVF_{MR}} = (1-\alpha \mathrm{MT_{sat}})\,(1-\nu_{iso})\,\nu_{icvf} \tag{4}$$

where $AWF = (1-\nu_{iso})\,\nu_{icvf}$ is the axonal water fraction estimated from the NODDI parameters (Stikov et al., 2015) and $MVF_{MR} = \alpha MT_{sat}$. The MR g-ratio was then computed according to Stikov et al. (2011, 2015)

$$\mathrm{g_{MR}} = \sqrt{1 - \frac{\mathrm{MVF_{MR}}}{\mathrm{MVF_{MR}} + \mathrm{AVF_{MR}}}} \tag{5}$$

Note that three versions of MT_{sat}, AVF_{MR}, and g_{MR} were generated according to notation in section "Data Processing": (i)

TABLE 1 | Group-averaged mean and standard deviation (SD) of g_{MR}^{B1}, MVF_{MR}^{B1}, and AVF_{MR}^{B1} in 21 high-SNR ROIs.

Name	Acronym	g_{MR}^{B1} mean ± SD	AVF_{MR}^{B1} mean ± SD	MVF_{MR}^{B1} mean ± SD
Anterior limb of internal capsule right	ACL r	0.688 ± 0.029	0.384 ± 0.052	0.419 ± 0.022
Retrolenticular part of internal capsule left	RIC l	0.665 ± 0.020	0.341 ± 0.025	0.428 ± 0.023
Anterior corona radiata right	ACR r	0.651 ± 0.012	0.321 ± 0.014	0.435 ± 0.014
Anterior corona radiata left	ACR l	0.644 ± 0.015	0.313 ± 0.014	0.440 ± 0.018
Superior corona radiata right	SCR r	0.679 ± 0.014	0.356 ± 0.018	0.413 ± 0.087
Superior corona radiata left	SCR l	0.674 ± 0.013	0.350 ± 0.016	0.419 ± 0.017
Genu of corpus callosum	GCC	0.642 ± 0.020	0.315 ± 0.021	0.445 ± 0.024
Body of corpus callosum	BCC	0.657 ± 0.021	0.328 ± 0.025	0.425 ± 0.020
Posterior corona radiata right	PCR r	0.662 ± 0.019	0.326 ± 0.025	0.416 ± 0.019
Posterior corona radiata left	PCR l	0.667 ± 0.018	0.337 ± 0.023	0.418 ± 0.019
Posterior thalamic radiation right	PTR r	0.643 ± 0.016	0.308 ± 0.017	0.438 ± 0.018
Posterior thalamic radiation left	PTR l	0.645 ± 0.017	0.313 ± 0.016	0.438 ± 0.020
Sagittal stratum left	SAS l	0.645 ± 0.021	0.314 ± 0.020	0.439 ± 0.025
External capsule right	EXC r	0.683 ± 0.020	0.359 ± 0.023	0.410 ± 0.028
External capsule left	EXC l	0.682 ± 0.025	0.357 ± 0.023	0.408 ± 0.034
Cingulum left	CGM l	0.661 ± 0.023	0.330 ± 0.028	0.422 ± 0.029
Fornix/Stria terminalis left	FNX l	0.669 ± 0.027	0.349 ± 0.036	0.426 ± 0.028
Superior longitudinal fasciculus right	SLF r	0.666 ± 0.016	0.334 ± 0.017	0.418 ± 0.022
Superior longitudinal fasciculus left	SLF l	0.668 ± 0.013	0.340 ± 0.015	0.420 ± 0.020
Superior fronto-occipital fasciculus right	SFO r	0.678 ± 0.020	0.361 ± 0.031	0.422 ± 0.020
Superior fronto-occipital fasciculus left	SFO l	0.672 ± 0.021	0.350 ± 0.029	0.424 ± 0.020

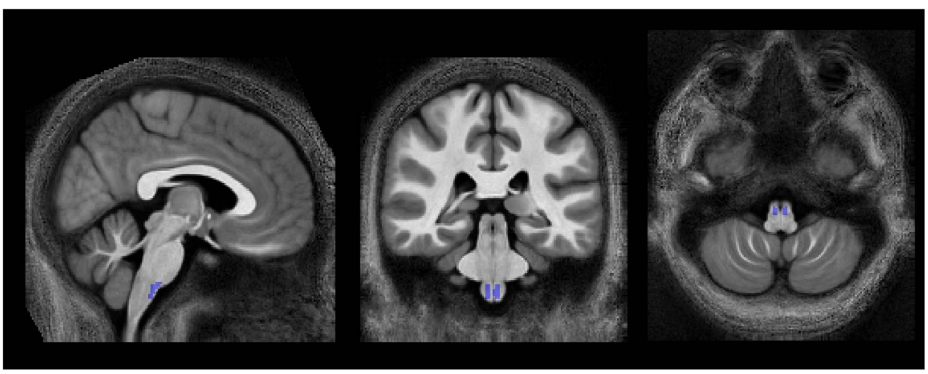


FIGURE 2 | Location of the pyramidal tracts in the medulla oblongata ROI, overlaid on the group-averaged MT_{sat}^{B1} map, that was used to determine the calibration constant, converting MT_{sat} into MVF_{MR} (section "Computation of MVF_{MR}, AVF_{MR}, and g_{MR}"). To create this ROI, the corticospinal tract ROI of the JHU-ICBM-DTI-81 atlas, which extends across the pons and medulla pyramids, was modified to cover only the medulla pyramids. Left-right position: X = 82; anterior-posterior position: Y = 77; superior-inferior position, Z = 30.

MVF_{MR}^{NO}, AVF_{MR}^{NO}, g_{MR}^{NO} for no correction, (ii) MVF_{MR}^{B1}, AVF_{MR}^{B1}, and g_{MR}^{B1} for B_1+ reference measurement, and (iii) MVF_{MR}^{UN}, AVF_{MR}^{UN}, and g_{MR}^{UN} for UNICORT B_1+ correction.

Definition of White Matter Masks

As g_{MR} and its constituents (MVF_{MR}, AVF_{MR}) are defined only in the WM, we restricted the analysis to the WM by creating binary WM masks (Mohammadi and Callaghan, 2020). WM tissue probability maps (WM-TPM) were created for each subject by segmenting AWF and MT_{sat}^{B1} using the hMRI toolbox, and taking their intersection according to Mohammadi and Callaghan (2020). In two subjects, the MT_{sat}^{B1} segmentation was of insufficient quality for segmentation and was replaced by the $R_1^{B_1}$ map. A group-specific binary WM mask (WM_{group}) was generated by averaging all individual WM-TPMs in the MNI space and thresholding it at 0.95.

A so-called high-SNR WM_{group} was also defined by taking the intersection of the WM_{group} and a binary signal-to-noise ratio (SNR) map. Hereby, the latter was used to reduce the number of voxels with unrealistically high values of v_{icvf} ($v_{icvf} \geq 0.999$). In 6 of 25 subjects, an SNR map was created by dividing the mean b_0 image by a single noise estimate in the native space and multiplied by the square root of the number of b_0 images per DWI dataset ($n = 12$). The noise was estimated within a noise ROI outside the brain in 72 images (6 subjects, both timepoints and 6 b0 images

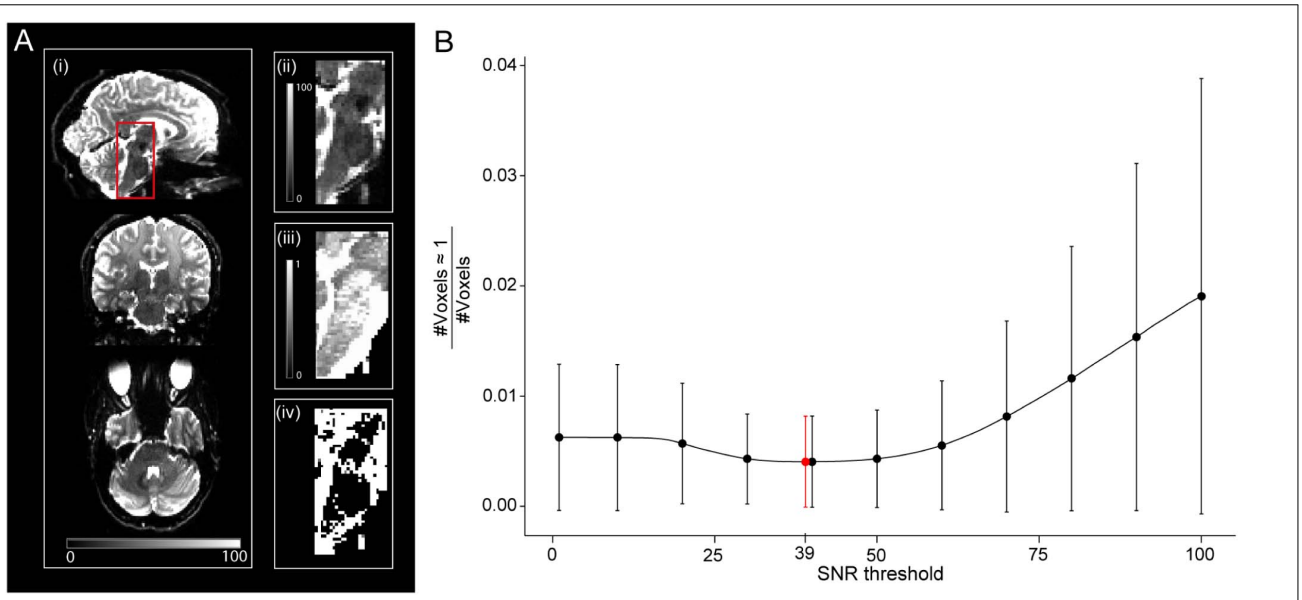


FIGURE 3 | Relationship between signal-to-noise ratio (SNR) and unrealistically high ν_{icvf} values—here defined as $\nu_{icvf} \geq 0.999$. **(A)** Sagittal, coronal, and axial view of the whole-brain SNR map (i), with a zoom-in view of the brainstem (ii). The brainstem is characterized by low SNR due to the spatial characteristics of the receive coil array (ii) and high occurrence of unrealistically high ν_{icvf} (iii), also shown as a binary mask (iv). **(B)** Given the co-occurrence of low SNR and unrealistically high ν_{icvf}, a binary SNR mask was created to exclude low-SNR voxels. To determine the optimal threshold for the SNR mask, the ratio between the number of voxels with unrealistically high ν_{icvf} and the total number of voxels within the mask were plotted against the SNR threshold. The solid dots and error bars represent the group mean and group standard deviation of the ratio, respectively. The SNR value that yielded the minimum of this ratio was considered optimal (SNR = 39, shown in red).

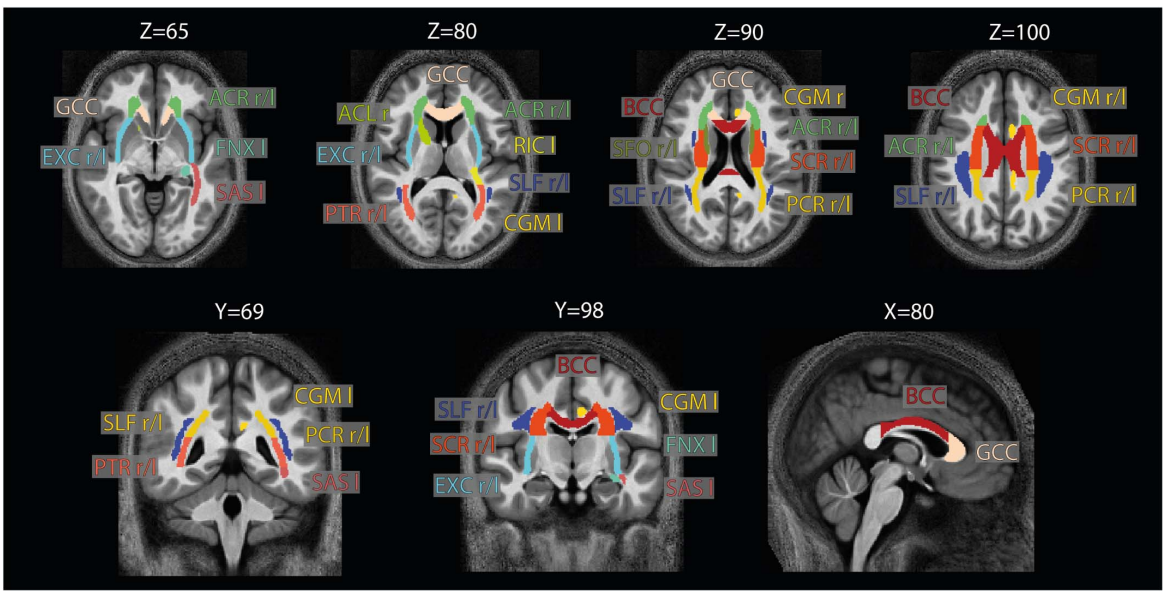


FIGURE 4 | Location of the ROIs used for analysis. The 21 high-SNR ROIs (listed in **Table 1**) are part of the JHU-ICBM-DTI-81 WM atlas (Hua et al., 2008) and are displayed here on the group-averaged normalized MT_{sat}^{B1} image. Note that for ROI analysis, the ROIs were projected into the native space using the inverse of the combined deformation field.

each) using the ACID toolbox, with the values averaged to obtain a single noise estimate. The threshold for SNR maps to create binary SNR map was chosen such that it minimizes the ratio between the number of artifactual voxels where $\nu_{icvf} \geq 0.999$ and the total number of voxels in the SNR mask (**Figure 3B**), yielding a value of 39. This was motivated by the observation that unrealistically high ν_{icvf} values typically occur in low-SNR areas (**Figures 3Aii,iii**). This threshold selection represents a trade-off between removing unrealistic voxels while retaining as many voxels as possible.

TABLE 2 | Summary statistics of g^{B1}_{MR}, AVF^{B1}_{MR}, and MVF^{B1}_{MR}.

	Δ_{DR}	$min_{i\in ROI}$	$max_{i\in ROI}$	mean	SD
g^{B1}_{MR},	0.046	0.642	0.688	0.664	0.014
AVF^{B1}_{MR}	0.076	0.308	0.384	0.337	0.020
MVF^{B1}_{MR}	0.037	0.408	0.445	0.425	0.010

This table lists the dynamic range (Δ_{DR}), lowest ($min_{i\in ROI}$) and highest ($max_{i\in ROI}$) ROI average value, mean value of the 21 analyzed ROI's (mean) with its corresponding standard deviation (SD).

Region of Interest Selection

For the region of interest (ROI) analysis, the JHU-ICBM-DTI-81 WM atlas (Hua et al., 2008) was transformed into the native space using the inverse of the combined deformation field. Two sets of ROIs were defined: (i) whole-WM ROIs and (ii) high-SNR ROIs, used for the main analysis. The whole-WM ROIs included those of the JHU-ICBM-DTI-81 WM atlas that were completely in WM_{group} defined in 2.6, yielding 43 ROIs (out of 48, leaving out the column and body of the fornix, the left and right cingulum part in the vicinity to the hippocampus, and the left and right uncinate fasciculus). The high-SNR ROIs included only those whole-WM ROIs that overlapped with the high-SNR WM_{group} to at least 95%, yielding 21 ROIs (**Figure 4** and **Table 2**). For the analyses, group-averaged g_{MR}, AVF_{MR}, and MVF_{MR} were calculated within the WM_{group}. Note that averaging included both sessions of each subject for all analyses except for the analysis in section "Test-Retest Analysis of the Group-Averaged MR G-ratio, Axon, and Myelin Volume Fraction."

Test-Retest Analysis of the Group-Averaged MR G-ratio, Axon, and Myelin Volume Fraction

The group-averaged g^{B1}_{MR} of the first and second session were compared within the previously mentioned 21 high-SNR ROIs using Bland-Altman plots (Bland and Altman, 1986). In the Bland-Altmann plots, the differences in g^{B1}_{MR} between the first ($(g^{B1}_{MR_1})$) and second ($(g^{B1}_{MR_2})$) session ($\delta^{retest}_i = (g^{B1}_{MR_1})_i - (g^{B1}_{MR_2})_i$) were plotted against their means ($mean^{retest}_i = \frac{(g^{B1}_{MR_1})_i + (g^{B1}_{MR_2})_i}{2}$), where i is the index of ROI i. Bias captures the offset ($\bar{\delta}^{retest} = \frac{1}{21}\sum_{i=1}^{21}\delta^{retest}_i$), while error ($\epsilon^{retest} = 1.96 \cdot \sqrt{\frac{1}{20}\sum_{i=1}^{21}(\delta^{retest}_i - \bar{\delta}^{retest})}$) captures the variation between the first and second scan within the i^{th} ROI. The computed $\bar{\delta}^{retest}$ and ϵ^{retest} were normalized by the dynamic range (Δ_{DR}) of g^{B1}_{MR} within the high-SNR ROIs, defined as $\Delta_{DR} = max_{i\in ROI}(mean^{retest}_i) - min_{i\in ROI}(mean^{retest}_i)$, yielding the relative error ($\delta^{retest}_{DR\%} = \frac{\epsilon^{retest}}{\Delta_{DR}} \cdot 100$) and relative bias ($\bar{\delta}^{retest}_{DR\%} = \frac{\bar{\delta}^{retest}}{\Delta_{DR}} \cdot 100$). *The same procedure was also applied to* AVF^{B1}_{MR} and MVF^{B1}_{MR}.

The distinction between bias and error is important, because while a potential bias can be retrospectively corrected, the error in the MR g-ratio method defines its sensitivity to detect differences between individuals, groups, or time points. To reliably capture these differences, the error must be significantly lower than the expected effect size.

Influence of B_1+ Correction in the Group-Averaged MR G-ratio, Axon, and Myelin Volume Fraction

Bland-Altman analysis was used to compare g_{MR} with and without B_1+ correction. In particular, the difference δ^{B1}_i in g_{MR} between $(g^{B1}_{MR})_i$, when using the reference method B_1+ correction, and $(g^{k}_{MR})_i$, when using no (k = NO) or UNICORT (k = UN) B_1+ correction: $\delta^{B1}_i = (g^{B1}_{MR})_i - (g^{k}_{MR})_i$ was plotted against their mean: $mean^{B1}_i = \frac{(g^{B1}_{MR})_i + \left(g^{k}_{MR}\right)_i}{2}$, with i being the index of the 21 high-SNR ROIs. The bias and error associated with the lack of (or UNICORT) B_1+ correction are defined as $\bar{\delta}^{B1} = \frac{1}{21}\sum_{i=1}^{21}\delta^{B1}_i$ and $\epsilon^{B1} = 1.96 \cdot \sqrt{\frac{1}{20}\sum_{i=1}^{21}(\delta^{B1}_i - \bar{\delta}^{B1})}$, respectively.

The computed ϵ^{B1} and $\bar{\delta}^{B1}$ were normalized by the dynamic range of g^{B1}_{MR} within the high-SNR ROIs, yielding the relative error ($\epsilon^{B1}_{DR\%} = \frac{\epsilon^{B1}}{\Delta_{DR}} \cdot 100$) and relative bias ($\bar{\delta}^{B1}_{DR\%} = \frac{\bar{\delta}^{B1}}{\Delta_{DR}} \cdot 100$). The same procedure was also applied to AVF_{MR} and MVF_{MR}, comparing them to their respective reference method and dynamic range. For MVF_{MR}, the Bland-Altman analysis was additionally done using the whole-WM ROIs instead of the high-SNR ROIs (see section "Region of Interest Selection") to assess the influence of including low-SNR voxels in the analysis.

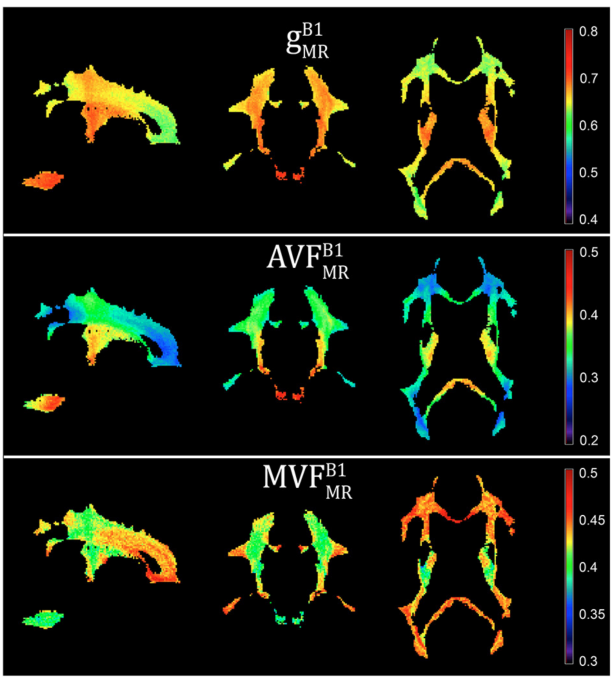


FIGURE 5 | Voxel-wise maps of group-averaged $g^{B1}{}_{MR}$, $AVF^{B1}{}_{MR}$, and $MVF^{B1}{}_{MR}$, restricted to the group WM mask (cf. section "Definition of White Matter Masks"). Depicted are a single sagittal (x = 100), coronal (y = 91), and axial (z = 85) slice.

Group Variability in MR G-ratio, Axon, and Myelin Volume Fraction

To assess group variability for each correction method, the coefficient-of-variation (CoV) across subjects and sessions was calculated for MVF_{MR}, AVF_{MR}, and g_{MR} in the MNI space after applying tissue-weighted smoothing (Tabelow et al., 2019), yielding: CoV^{B1}_{MR}, CoV^{UN}_{MR}, and CoV^{NO}_{MR}, where MR $\in$ {g_{MR}, AVF_{MR}, and MVF_{MR}}. For tissue-weighted smoothing, a full width at half maximum Gaussian smoothing kernel of 6 mm was used. Bland-Altman analysis (see section "Test-Retest Analysis of the Group-Averaged MR G-ratio, Axon, and Myelin Volume Fraction") was used to compare CoV^{UN}_{MR} and CoV^{NO}_{MR} against CoV^{B1}_{MR} based on the reference method, yielding bias ($\bar{\delta}^{CoV}$) and error (ϵ^{CoV}) values. A higher variability across the brain is expected to increase $\bar{\delta}^{CoV}$ whereas a higher local variability is expected to increase ϵ^{CoV}.

RESULTS

G-ratio, Myelin, and Axonal Volume Fraction Across the White Matter

Voxel-wise maps of group-averaged g^{B1}_{MR}, AVF^{B1}_{MR}, and MVF^{B1}_{MR} in WM are shown in **Figure 5**. The group-averaged mean and standard deviation of g^{B1}_{MR}, MVF^{B1}_{MR}, and AVF^{B1}_{MR} in 21 high-SNR ROIs are reported in **Table 1** and **Figure 6**. The dynamic range (Δ_{DR}), minimum and maximum values, and mean and standard deviation of g^{B1}_{MR}, AVF^{B1}_{MR}, and MVF^{B1}_{MR} across ROIs are listed in **Table 2**. The largest g^{B1}_{MR} and AVF^{B1}_{MR} were found in the right anterior limb of the internal capsule (0.688 and 0.384, respectively), while the largest MVF^{B1}_{MR} was in the genu of corpus callosum (0.445), where also the lowest g^{B1}_{MR} (0.642) can be found. The lowest AVF^{B1}_{MR}, and MVF^{B1}_{MR} were found in the right posterior thalamic radiation (AVF^{B1}_{MR} = 0.308) and in the left external capsule (MVF^{B1}_{MR} = 0.408), respectively. The Δ_{DR} was the smallest for MVF^{B1}_{MR} (0.037), followed by g^{B1}_{MR} (0.046) and AVF^{B1}_{MR} (0.076).

Test-Retest Analysis of the Group-Averaged MR G-ratio, Axon, and Myelin Volume Fraction

The relative error ($\epsilon^{retest}_{DR\%}$) and bias ($\bar{\delta}^{retest}_{DR\%}$) values of the test-retest analysis are summarized in **Table 3** and shown as Bland-Altmann plots in **Figure 7**. The test-retest analysis

TABLE 3 | Bias and error between scans, in g^{B1}_{MR}, AVF^{B1}_{MR}, and MVF^{B1}_{MR}.

MAP	$\bar{\delta}^{retest}$	ϵ^{retest}	$\bar{\delta}^{retest}_{DR\%}$	$\epsilon^{retest}_{DR\%}$
g^{B1}_{MR}	0.0021	0.0102	4.57	22.17
AVF^{B1}_{MR}	0.0006	0.0156	0.79	20.53
MVF^{B1}_{MR}	−0.0031	0.0076	−8.38	20.54

*List of the bias ($\bar{\delta}^{retest}$) and error (ϵ^{retest}) values, defined as in **Figure 7**, along with their relative value with respect to the dynamic range Δ_{DR}: $\epsilon^{retest}_{DR\%} = \frac{\epsilon^{retest}}{\Delta_{DR}} \cdot 100$; $\bar{\delta}^{retest}_{DR\%} = \frac{\bar{\delta}^{retest}}{\Delta_{DR}} \cdot 100$.*

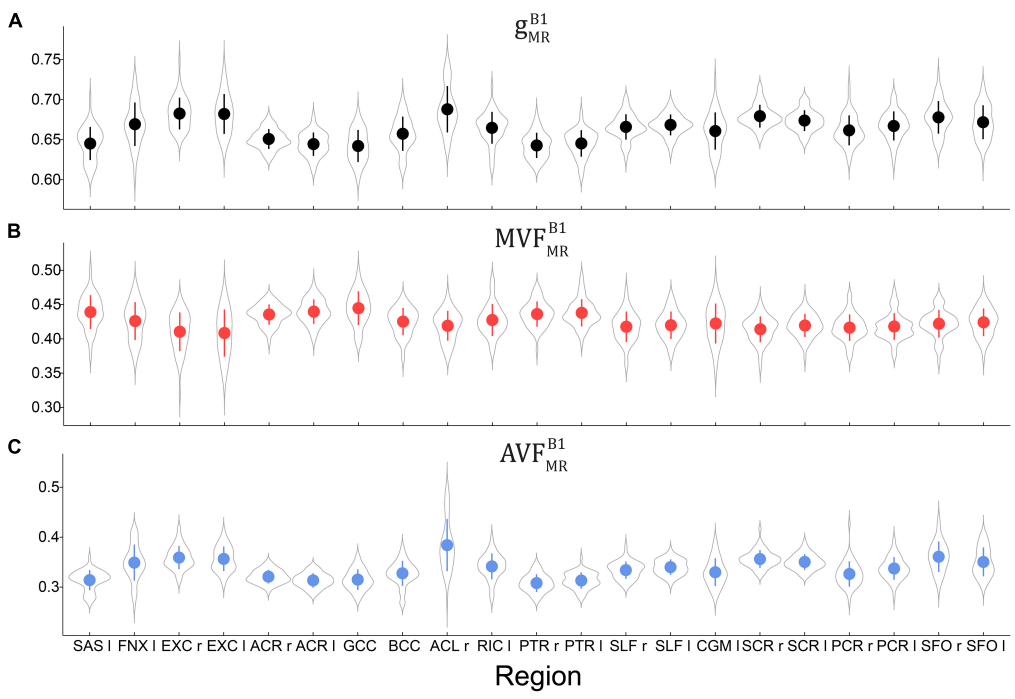


FIGURE 6 | Violin plots representing the distribution of g^{B1}_{MR} **(A)**, MVF^{B1}_{MR} **(B)**, and AVF^{B1}_{MR} **(C)** across the group and in 21 high-SNR ROIs listed in **Table 1**. The mean and standard deviation of the distribution are indicated by solid dot and whiskers, respectively.

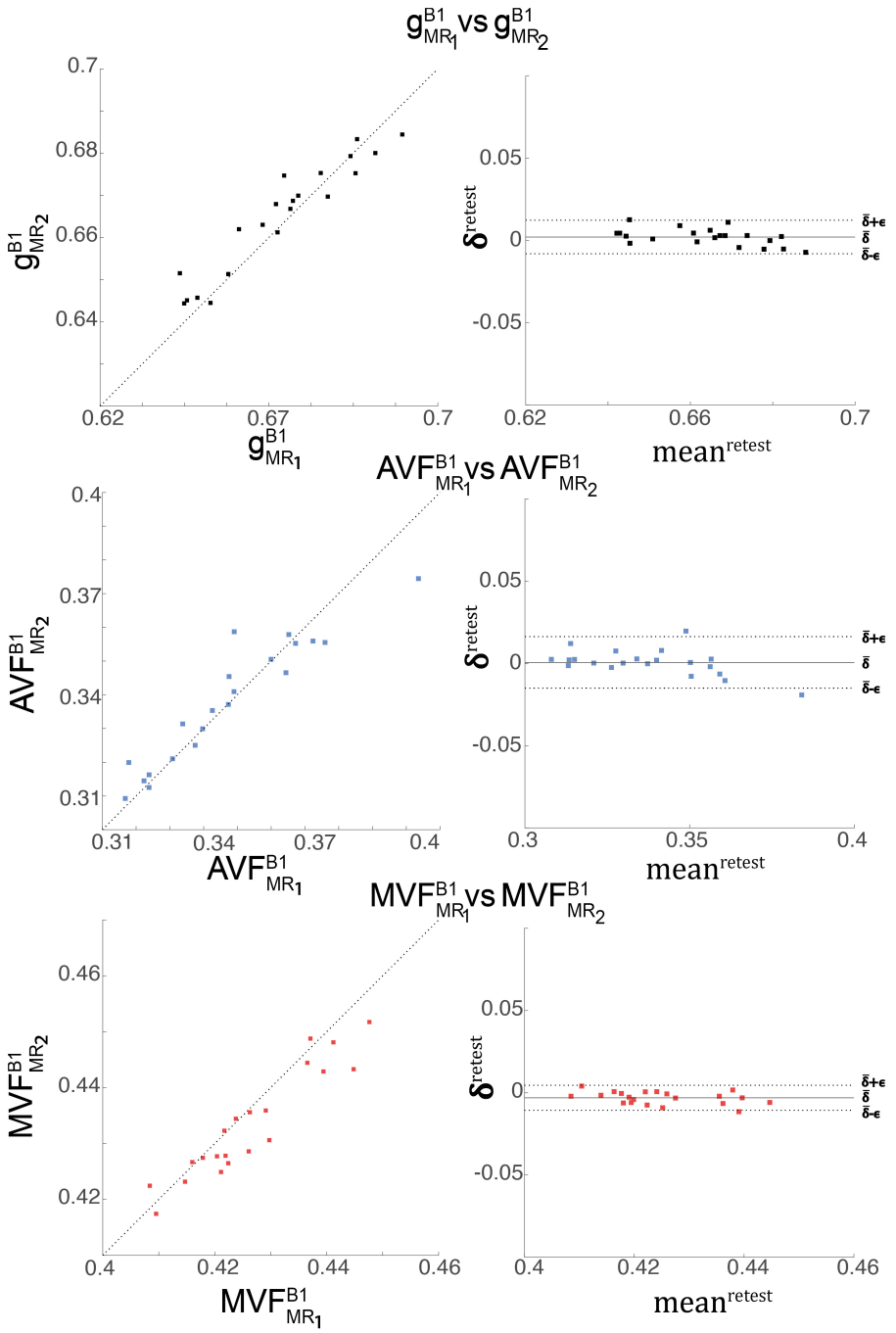


FIGURE 7 | Depicted are scatter and Bland-Altman plots of g^{B1}_{MR} (first row), AVF^{B1}_{MR}(second row), and MVF^{B1}_{MR} (third row) from two session across 21 WM regions (denoted high-SNR ROIs, see **Figure 4**). The Bland-Altman plot illustrates the differences between values obtained from the two sessions (e.g., $g^{B1}_{MR_1}$ vs. g^{B1}_{MR2}; $\delta_i^{retest} = (g^{B1}_{MR1})_i - (g^{B1}_{MR2})_i$) against their mean (e.g., $mean_i^{retest} = \frac{(g^{B1}_{MR1})_i + (g^{B1}_{MR2})_i}{2}$, with *i* indexing the i^{th} ROI). Each point in the scatter plot represents the group-averaged value in a single ROI. The bold black line represents the bias ($\bar{\delta}^{retest} = \frac{1}{21}\sum_{i=1}^{21}\delta_i^{retest}$), while the dashed line shows error ($\epsilon^{retest} = 1.96 \cdot SD(\delta_i^{retest})$) between the two sessions.

revealed a $\bar{\delta}^{retest}_{DR\%}$ below an absolute value of 8.4% for each metric (g^{B1}_{MR}, AVF^{B1}_{MR}, and MVF^{B1}_{MR}), where the AVF^{B1}_{MR} showed the lowest $\bar{\delta}^{retest}_{DR\%}$ with 0.79% (**Figure 7** and **Table 3**). The $\epsilon^{retest}_{DR\%}$ was below 22.2% for each metric, where the AVF^{B1}_{MR} showed the lowest $\epsilon^{retest}_{DR\%}$ with 20.5% (**Figure 7** and **Table 3**).

TABLE 4 | Bias and error between methods, in g_{MR}, AVF_{MR}, and MVF_{MR}.

MAP	$\bar{\delta}^{B1}$	ϵ^{B1}	$\bar{\delta}^{B1}_{DR\%}$	$\epsilon^{B1}_{DR\%}$
g^{B1}_{MR} vs. g^{NO}_{MR}	−0.041	0.017	−89.13	36.96
g^{B1}_{MR} vs. g^{UN}_{MR}	0.014	0.005	30.44	10.87
AVF^{B1}_{MR} vs. AVF^{NO}_{MR}	−0.031	0.012	−40.79	15.79
AVF^{B1}_{MR} vs. AVF^{UN}_{MR}	0.011	0.004	14.47	5.26
MVF^{B1}_{MR} vs. MVF^{NO}_{MR}	0.053	0.022	143.24	59.46
MVF^{B1}_{MR} vs. MVF^{UN}_{MR}	−0.018	0.006	−48.65	16.22
EWM MVF^{B1}_{MR} vs. MVF^{NO}_{MR}	0.033	0.048	36.48	52.75
EWM MVF^{B1}_{MR} vs. MVF^{UN}_{MR}	−0.012	0.022	−13.08	23.96

*List of the bias ($\bar{\delta}^{B1}$) and error (ϵ^{B1}) values as defined in **Figure 9**, along with their relative value with respect to the dynamic range Δ_{DR}: $\epsilon^{B1}_{DR\%} = \frac{\epsilon^{B1}}{\Delta_{DR}} \cdot 100$; $\bar{\delta}^{B1}_{DR\%} = \frac{\bar{\delta}^{B1}}{\Delta_{DR}} \cdot 100$. Note that the error and bias in the last two rows were obtained when using the whole-WM ROIs instead of the high-SNR ROIs (see **Supplementary Figure 1**).*

Influence of B_1+ Correction on the Group-Averaged MR G-ratio, Axon, and Myelin Volume Fraction

The relative error ($\epsilon^{B1}_{DR\%}$) and bias ($\bar{\delta}^{B1}_{DR\%}$) values of the B_1+ correction analysis are summarized in **Table 4** and shown as Bland-Altmann plots in **Figures 8**, **9**. For g_{MR}, compared to the no-correction case, UNICORT showed both lower $\epsilon^{B1}_{DR\%}$ (UNICORT vs. no correction: 10.9% vs. 37.0%) and $\bar{\delta}^{B1}_{DR\%}$ (30.4% vs. −89.1%). For both AVF_{MR} and MVF_{MR}, UNICORT yielded lower $\epsilon^{B1}_{DR\%}$ (UNICORT vs. no correction; AVF_{MR}: 5.3% vs. 15.8%; 16.2% vs. 59.5%) and lower $\bar{\delta}^{B1}_{DR\%}$ (AVF_{MR}: 14.5% vs. −40.8%; MVF_{MR}: 48.6% vs. 143.2%). Altogether, the UNICORT correction reduced the bias and error in the MR g-ratio and its constituents by roughly a factor of three. The lower $\epsilon^{B1}_{DR\%}$ and $\bar{\delta}^{B1}_{DR\%}$ associated with UNICORT was also reflected by the fact that values of g^{UN}_{MR}, AVF^{UN}_{MR}, and MVF^{UN}_{MR} (**Figure 8**, lower panel) lie closer to the unit slope line than values of g^{NO}_{MR}, AVF^{NO}_{MR}, and MVF^{NO}_{MR} (**Figure 8**, upper panel). When computing $\epsilon^{B1}_{DR\%}$ and $\bar{\delta}^{B1}_{DR\%}$ of g_{MR} in the whole-WM ROIs (see **Supplementary Figure 1**), $\bar{\delta}^{B1}_{DR\%}$ was consistently lower for both the no-correction case (whole-WM ROIs vs. high-SNR ROIs: 36.5% vs. 143.2%) and UNICORT (13.1% vs. 48.6%), whereas $\epsilon^{B1}_{DR\%}$ was similar (no-correction: 52.8% vs. 59.5%; UNICORT: 24.0% vs. 16.2%).

Group Variability in MR G-ratio, Axon, and Myelin Volume Fraction

g_{MR} showed on average smaller CoV than AVF_{MR} and MVF_{MR} (**Figure 10**). In all maps, the CoV was the highest in the deep brain areas. The relative error ($\frac{\epsilon^{CoV}}{CoV^{B1}} \cdot 100$) and bias ($\frac{\bar{\delta}^{CoV}}{CoV^{B1}} \cdot 100$) values of CoV with respect to the B_1+ reference measurement

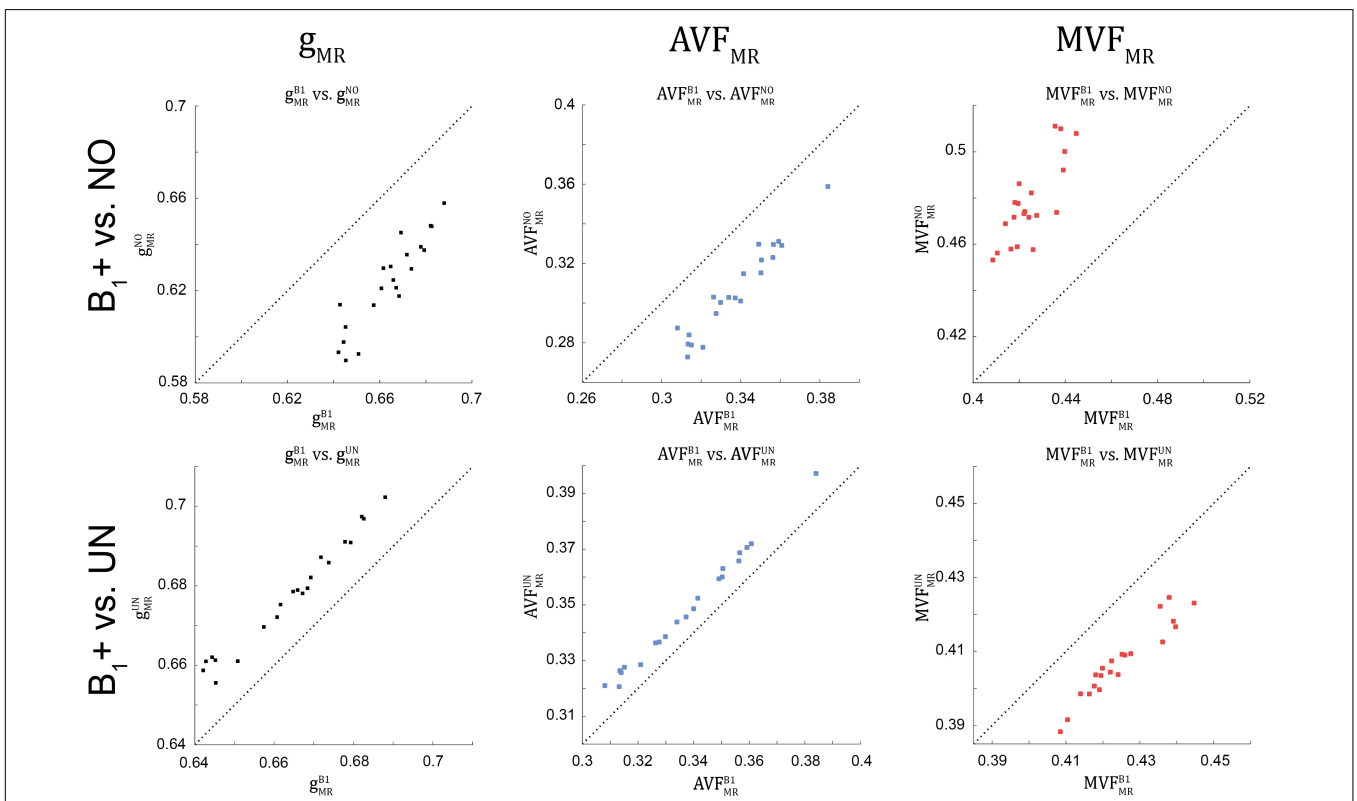


FIGURE 8 | Scatter plots of g_{MR}, AVF_{MR}, and MVF_{MR}, plotting values obtained without B_1+ correction (superscript: NO, top row) and with UNICORT B_1+ correction (superscript: UN, bottom row) against values obtained with the reference method, i.e., B_1+ field map correction (superscript: B1). A dashed unit slope line is plotted for reference. Each point in the scatter plot represents the group-averaged value in a single ROI (see **Figure 4** for the locations of the 21 high-SNR ROIs).

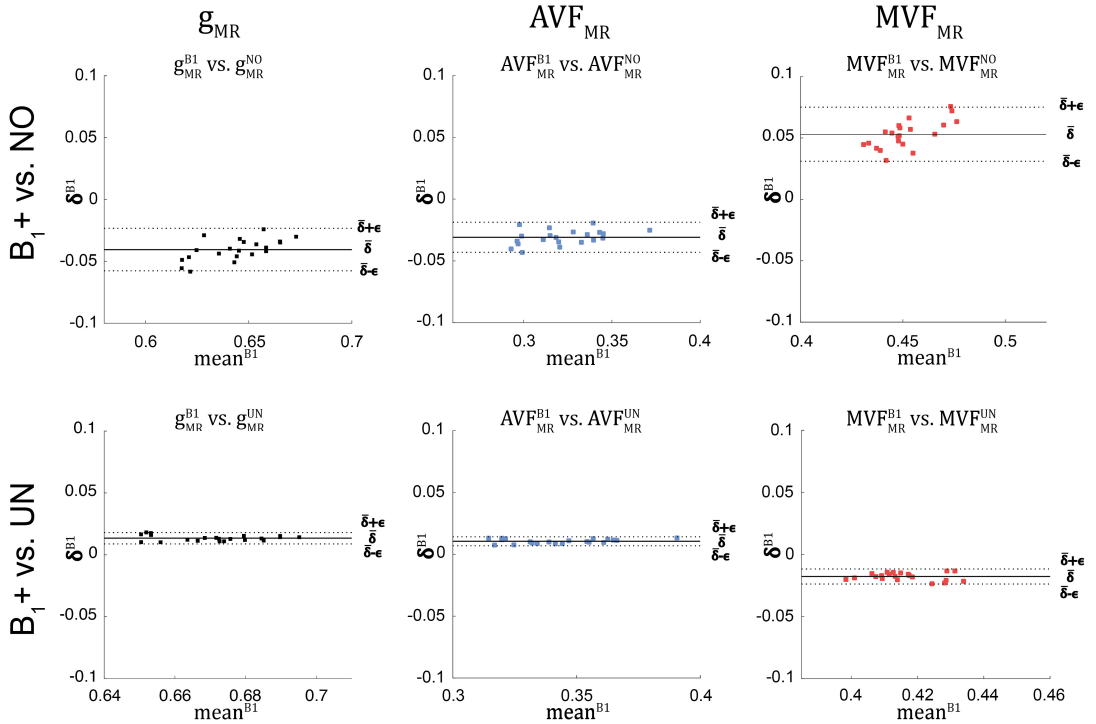


FIGURE 9 | Bland-Altman plots of g_{MR}, AVF_{MR}, and MVF_{MR}, comparing values obtained without B_1+ correction (NO, top row) and with UNICORT B_1+ correction (UN, bottom row) against values obtained by B_1+ field map correction (superscript: B1). The Bland-Altman plot illustrates the differences between values obtained by two different methods (reference vs. tested method); e.g., $\delta_i^{B1} = (g_{MR}^{B1})_i - (g_{MR}^{k})_i$ against their mean ($mean_i^{B1} = \frac{(g_{MR}^{B1})_i + (g_{MR}^{k})_i}{2}$, with $k = \{UN, NO\}$ and i indexing the i^{th} ROI). Each point in the scatter plot represents the group-averaged value in a single ROI (see **Figure 4** for the locations of the 21 high-SNR ROIs). The bold black line represents the bias ($\bar{\delta}^{B1} = \sum_{i=1}^{21} \delta_i^{B1}$), while the dashed line shows error ($\epsilon^{B1} = 1.96 \cdot SD(\delta_i^{B1})$) between the reference and the tested method. Error and bias values averaged across all ROIs and subjects are listed in **Table 5**.

are summarized in **Table 5** and the error and bias are also displayed as Bland-Altman density plot in **Figure 11**. For g_{MR}, compared to the no correction case, UNICORT showed similar ϵ^{CoV} (UNICORT vs. no correction: 0.6% vs. 0.6%) but lower $\bar{\delta}^{CoV}$ (−0.1% vs. −0.4%). UNICORT yielded higher ϵ^{CoV} (UNICORT vs. no correction; 1.0% vs. 0.8%) and lower $\bar{\delta}^{CoV}$ (−0.2% vs. −0.4%) for AVF_{MR}, and higher ϵ^{CoV} (1.2% vs. 0.4%) and higher $\bar{\delta}^{CoV}$ (−0.5% vs. −0.1%) for MVF_{MR}. The lower $\bar{\delta}^{CoV}$ of g_{MR} and AVF_{MR} associated with UNICORT reveals itself as a slight shift of the points toward the unit slope line in the scatter density plot (**Figure 12**).

DISCUSSION

In this study, we showed that omitting the correction of the magnetization transfer saturation map (MT_{sat}) for residual B_1+ effects introduces large error and bias in the MR g-ratio and the constituents (myelin and axon volume fractions, or in short MVF_{MR} and AVF_{MR}). We also demonstrated that this error and bias can be reduced by roughly a factor of three using the data-driven UNICORT B_1+ correction (implemented in the hMRI toolbox, see text footnote 1) when a B_1+ field measurement is unavailable.

The Effect of Omitting the B_1+ Field Measurement

MT_{sat} have been often used as a proxy for the MVF_{MR} in g-ratio weighted imaging (Mohammadi et al., 2015; Campbell et al., 2018; Ellerbrock and Mohammadi, 2018; Hori et al., 2018; Kamagata et al., 2019), because they are directly linked to the macromolecular pool with an intrinsic correction for underlying longitudinal relaxation time and B_1+ field inhomogeneities effects (Helms et al., 2008). Despite the latter intrinsic correction for B_1+ field inhomogeneities, we found that the residual B_1+ effects on MT_{sat} map were still observable. In particular, the bias and error of the MR g-ratio (g_{MR}) was about −89 and 37% higher, respectively, when omitting the B_1+ correction. We found the same trend for MVF_{MR} and AVF_{MR}; while the error and bias were even larger for MVF_{MR} when B_1+ correction was omitted, it was smaller but still substantial for the AVF_{MR}. We found that omitting B_1+ leads to a substantially higher (more than 10-fold) bias in the MR g-ratio and its constituents when compared to a test-retest analysis of our data (**Figure 7** and **Table 3**). Also, the error due to omitting the B1+ correction was twice as large as the error observed in the test retest analysis for the MR g-ratio and the MVF, whereas for AVF the errors were similar. We expect that the high error will be of particular

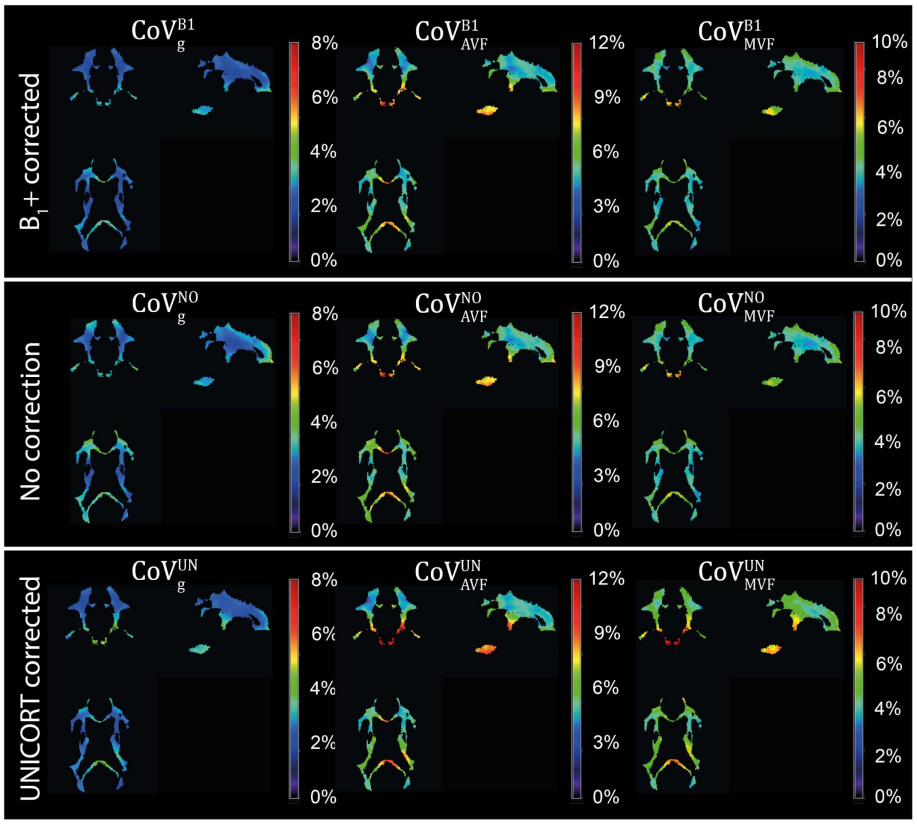


FIGURE 10 | Coefficient of variation (CoV) maps of g_{MR}, AVF_{MR}, and MVF_{MR} with B_1+ correction (CoV_g^{B1}, CoV_{AVF}^{B1}, and CoV_{MVF}^{B1}), no correction (CoV_g^{NO}, CoV_{AVF}^{NO}, and CoV_{MVF}^{NO}), and UNICORT B_1+ correction (CoV_g^{UN}, CoV_{AVF}^{UN}, and CoV_{MVF}^{UN}). CoV maps, expressed in percentage, were computed as the voxel-wise ratio between the group mean and group standard deviation maps of the normalized g_{MR}, AVF_{MR}, or MVF_{MR}. The voxel-wise computation of CoV is restricted to the group WM mask (cf. section "Definition of White Matter Masks"). Shown are a single coronal (y = 91), sagittal (x = 100), and axial (z = 85) slice.

relevance for group studies because it can be regarded as an error that evolves when replacing the reference method with the alternative method. For comparison, age-related changes assessed by g-ratio weighted imaging (Cercignani et al., 2017; Berman et al., 2018) have been reported to vary between 30 and 100% (in absolute values: g_{MR}0.02–0.04 (**Figure 5** in Cercignani et al., 2017). Consequently, the reported effect size of age-related changes would have become potentially undetectable if the B_1+ field correction has been omitted in the study of Cercignani et al. (2017). The B_1+ effect is particularly relevant for the MR g-ratio method by Cercignani et al. (2017) that combined quantitative MT (Gloor et al., 2008) with NODDI, because the qMT method does not possess an intrinsic correction for B_1+ field inhomogeneities as opposed to the MT_{sat} methods used here. Note that we reported, for better intuition, the bias and error relative to the dynamic range of the parameters across the investigated white matter (WM) ROIs (the dynamic range of g_{MR} is $\Delta_{DR} = 0.046$; the absolute bias and error can be found in **Table 4**).

To reduce this source of bias and error, we propose a data-driven approach to correct for B_1+ field inhomogeneities when no B_1+ field measurement is available. To this end, we used UNICORT to estimate the B_1+ field (Weiskopf et al., 2011). We found that using the UNICORT-estimated B_1+ field to correct residual B_1+ field inhomogeneities in MT_{sat} reduces at the group level the bias and error in the MR g-ratio and its constituents by roughly a factor of three. However, the UNICORT estimated B_1+ inhomogeneity can be erroneous with the error varying across subjects. To assess this variability, we estimated coefficient-of-variance (CoV) maps of g_{MR}, AVF_{MR}, and MVF_{MR} for all

TABLE 5 | Bias and error between methods, in the CoV of g_{MR}, AVF_{MR}, and MVF_{MR}.

MAP	$\bar{\delta}^{CoV}$	ϵ^{CoV}	$\frac{\bar{\delta}^{CoV}}{CoV_{MR}^{B1}} \cdot 100$	$\frac{\epsilon^{CoV}}{CoV_{MR}^{B1}} \cdot 100$
CoV g_{MR}^{B1} vs. CoV g_{MR}^{NO}	−0.42	0.56	−17.3	23.1
CoV g_{MR}^{B1} vs. CoV g_{MR}^{UN}	−0.12	0.62	−4.9	25.5
CoV AVF_{MR}^{B1} vs. CoV AVF_{MR}^{NO}	−0.40	0.78	−7.3	14.3
CoV AVF_{MR}^{B1} vs. CoV AVF_{MR}^{UN}	−0.21	1.02	−3.8	18.7
CoV MVF_{MR}^{B1} vs. CoV MVF_{MR}^{NO}	−0.05	0.41	−1.1	9.2
CoV MVF_{MR}^{B1} vs. CoV MVF_{MR}^{UN}	−0.52	1.20	−11.9	27.0

*List of the bias ($\bar{\delta}^{CoV}$) and error (ϵ^{CoV}) values as defined in **Figure 11**, along with their relative value with respect to the group-average CoV across the MR g-ratios using the reference B_1+ field correction method: $\frac{\bar{\delta}^{CoV}}{CoV_{MR}^{B1}} \cdot 100$; $\frac{\epsilon^{CoV}}{CoV_{MR}^{B1}} \cdot 100$.*

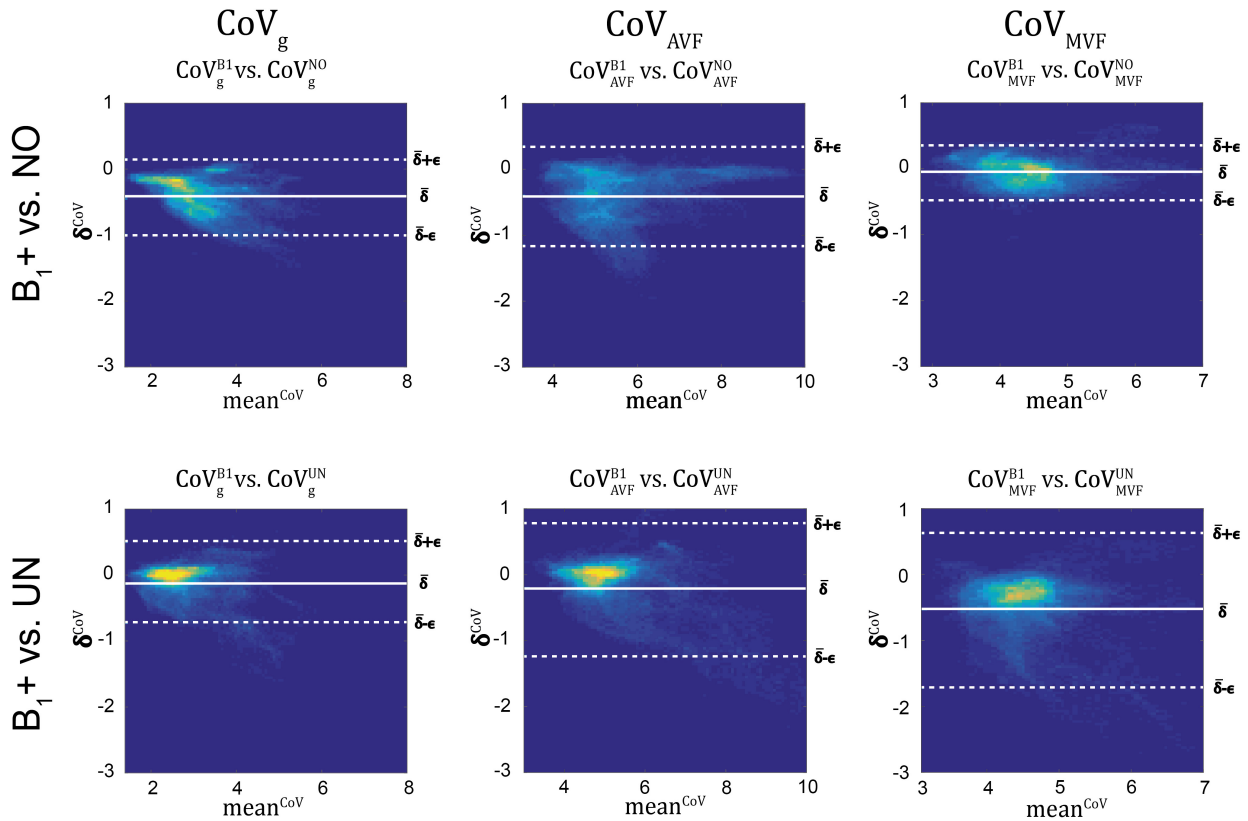


FIGURE 11 | Bland-Altman density plots of CoV_g, CoV_{AVF}, and CoV_{MVF} for no correction (NO, top row) and UNICORT B_1+ correction (UN, bottom row) against the reference method (B_1+) (yellow indicates high density and blue low). The Bland-Altman plot depicts the differences between the tested parameter maps and the reference method (e.g., $\delta_i^{CoV} = (CoV_g^{B1})_i - (CoV_g^k)_i$) against their mean (e.g., $mean_i^{CoV} = \frac{(CoV_g^{B1})_i + CoV_g^k)_i}{2}$) with k = {UN, NO} and *i* being the index of the i^{th} region. The bold white line represents the bias ($\bar{\delta}^{CoV} = \sum_{i=1}^{N} {}_i^{CoV}$; N = number of voxels) and the dashed lines represent $\bar{\delta}^{CoV}$ the error ($\epsilon^{CoV} = 1.96 \cdot SD(\delta_i^{CoV})$)]. The error and bias values are summarized in **Table 5**.

three methods. In general, an increased CoV can be found at tissue boundaries (e.g., cerebral spinal fluid to WM) due to slight misregistration between the maps of axonal and myelin markers and/or imperfect normalization (**Figure 10**). Additionally, we found a strong increase in the bias and error of the CoV of MVF maps (increase in bias: 11% and in error: 18%) when UNICORT B_1+ correction was used as compared to no correction. The CoV of g_{MR} and AVF_{MR} did not show a consistent trend: while the bias decreased, the error increased for both parameters. In other words, the UNICORT B_1+ correction leads to higher accuracy in the g-ratio and its constituents but comes at the cost of a lower precision in MVF.

G-ratio, Myelin, and Axonal Volume Fraction Across the White Matter

Our g_{MR}^{B1} and AVF_{MR}^{B1} across the white matter were within the range of the reported values of previous studies (g_{MR}: 0.64–0.76; AVF_{MR}: 0.26–0.43 in (Cercignani et al., 2017; Berman et al., 2018). The range of MVF_{MR}^{B1} was in the upper half of previously reported values (0.17–0.42 in Cercignani et al., 2017). Our slightly higher MVF_{MR} values might be due to differences in the calibration approach: while we calculated the reference MVF_{REF} from previously published *ex-vivo* histology data (Graf von Keyserlingk and Schramm, 1984), Cercignani et al. (2017), used a reference from previously published *ex-vivo* histology g-ratio data in the corpus callosum and Berman et al. (2018), did not perform any calibration assuming that macromolecular tissue volume and MVF_{MR} are equal. An error in the calibration constant can lead to a bias in the MVF estimates which in turn leads to an error and bias in the MR g-ratio (Campbell et al., 2018).

Confounding Factors

As this study calculates the *in-vivo* MR g-ratio, there is no histological data available from the participants of this study, which could be used for calibration or as a gold standard reference. For calibration of MT_{sat} to MVF_{MR}, we estimated the histological MVF (MVF_{hist}) from published *ex-vivo* data within the human medulla oblongata (Graf von Keyserlingk and Schramm, 1984). Since the reference MVF_{hist} and the calibrated MT_{sat} map were taken from different subjects, this might introduce a systematic bias in the MR g-ratio. However, since we found a relatively good agreement between our g_{MR},

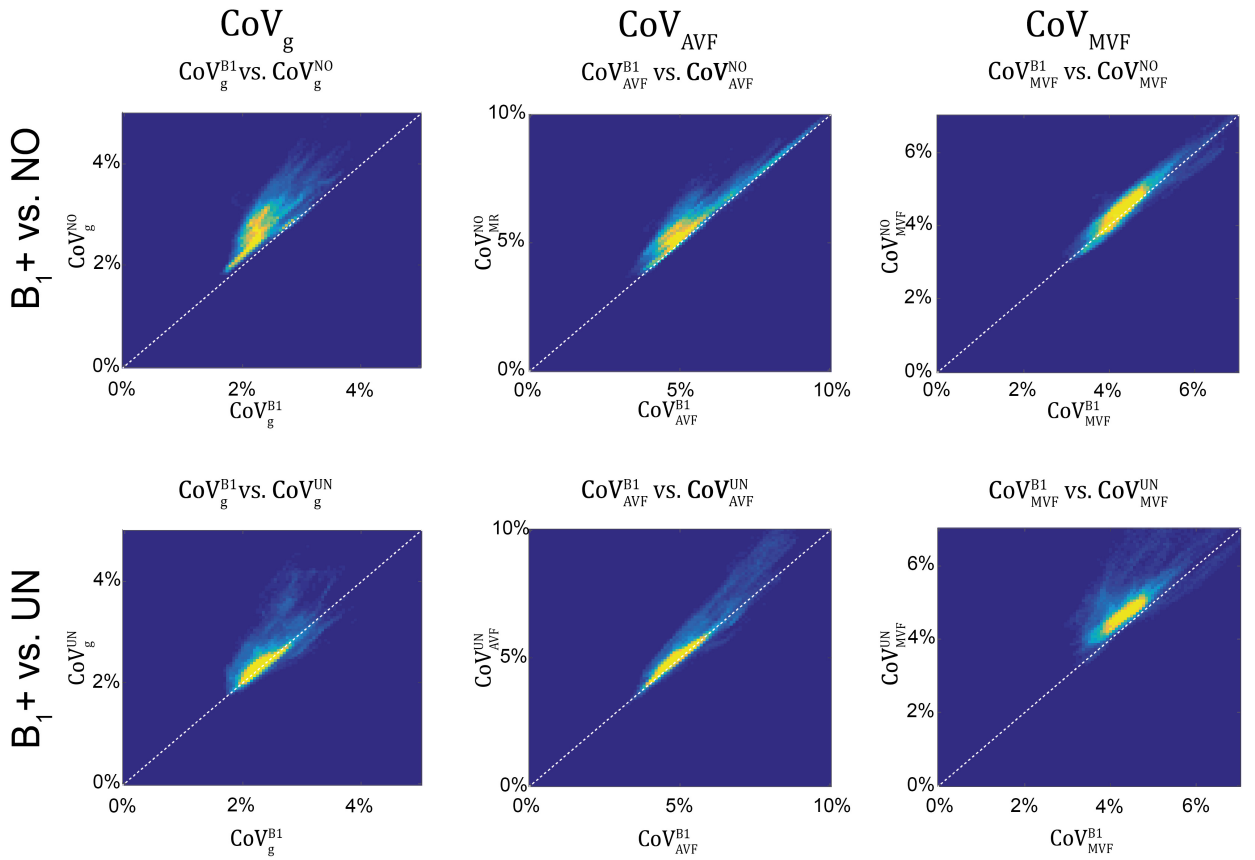


FIGURE 12 | Scatter density plots of CoV_g (left column), CoV_{AVF} (middle column), and CoV_{MVF} (right column), plotting values obtained with no correction (superscript: NO, top row) and with UNICORT B_1+ correction (UN, bottom row) against values obtained by B_1+ field map correction (superscript: B1). The unit slope line is plotted for orientation (dotted line). The dots in the scatter plots represent the WM voxels in the CoV maps in **Figure 10** (yellow indicates high voxel density).

AVF_{MR}, and MVF_{MR} values with previously reported values obtained by a different calibration approach (Cercignani et al., 2017; Berman et al., 2018), we expect that it had a small effect on the results. Moreover, we focused on the effect of omitting B_1+ correction, which will lead to additional inaccuracies in g-ratio weighted imaging, independent of the quality of the calibration.

Although, not reported in previous NODDI-based g-ratio mapping studies (Stikov et al., 2015; Cercignani et al., 2017; Jung et al., 2017; Mancini et al., 2017; Ellerbrock and Mohammadi, 2018; Hori et al., 2018), we found that the intra-cellular volume fraction (ν_{icvf}) determined with NODDI tends to be biased at small signal-to-noise ratios (SNR $<$ 39), resulting in a ceiling effect, i.e., $\nu_{icvf} \approx 1$. To avoid a corresponding bias in g_{MR} (and AVF_{MR}), we restricted the analysis to regions with sufficiently high SNR (**Figure 3**). To investigate whether our findings generalize to low-SNR regions as well, we performed an additional Bland-Altman analysis of MVF_{MR} in whole-WM ROIs. To this end, a larger set of ROIs was used covering the entire white matter. Although the bias was smaller for the whole-WM as compared to the high-SNR ROI analysis, we found the same trend: the error and bias were reduced when using UNICORT B_1+ correction relative to no correction. Note that the smaller bias for the whole-WM analysis is most probably an artifact of the calibration procedure. Since the ROI used for calibration was not part of the high-SNR ROIs but was part of the whole-WM ROIs, we think it could have reduced the bias in the whole-WM ROI analysis as compared to the high-SNR analysis.

We note that the presented results were based on a customized B_1+ mapping method (Lutti et al., 2010). Using vendor specific protocols for B_1+ and MT_{sat} mapping may influence the results (Leutritz et al., 2020). Moreover, the calibration factor in Equation (1) may have to be recalibrated for different MT-pulses.

Future studies should investigate the effect of B_1+ correction on MR g-ratio mapping when using alternative biomarkers to estimate AVF_{MR} and MVF_{MR} (e.g., Ellerbrock and Mohammadi, 2018). Moreover, there are alternative B_1+ mapping approaches available which might vary in precision (Lutti et al., 2010) and therefore can affect the MR g-ratio values. However, the differences in the precision of these methods are in the order of few percentage and thus much smaller than the effect of omitting the B_1+ field or using the data-driven UNICORT B_1+ estimate (Weiskopf et al., 2011).

CONCLUSION

In this study, we assessed the effect of B_1+ correction on the accuracy of MR g-ratio as well as axonal and myelin volume fraction based on MT_{sat} and NODDI. Our results demonstrate that B_1+ correction via a measured B_1+ field map is the method of choice. If the B_1+ field map cannot be acquired, we propose the retrospective, data-driven UNICORT B_1+ correction to estimate and correct for B_1+ field inhomogeneities, which reduces the error and bias by a factor of three. UNICORT is implemented in the free and open-source hMRI toolbox (see text footnote 1).

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data that support the findings of this study are available on request from the corresponding author. The data have not been made freely available on the internet due to privacy or ethical restrictions. Requests to access the datasets should be directed to corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ärztekammer Hamburg. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SM and TE contributed to the conception and design of the study, performed statistical analysis and MRI processing, and wrote the first draft of the manuscript. All authors contributed substantially to revising the manuscript critically for intellectual content and have approved the submitted version.

FUNDING

This project was by the ERA-NET NEURON (hMRIofSCI), the Federal Ministry of Education and Research (BMBF; 01EW1711A and B), and the German Research Foundation (DFG Priority Program 2041 "Computational Connectomics," (AL 1156/2-1;GE 2967/1-1; MO 2397/5-1; MO 2249/3–1), DFG Emmy Noether Stipend: MO 2397/4-1), and the Forschungszentrums Medizintechnik Hamburg (fmthh; grant 01fmthh2017). GH was supported by the Swedish Research Council (NT 2014-6193). PF was funded by a SNF Eccellenza Professorial Fellowship grant (PCEFP3_181362/1), and the European Union's Horizon 2020 research, innovation programme (grant agreement no 634541). EB received funding from the European Structural and Investment Fund/European Regional Development Fund and the Belgian Walloon Government, project BIOMED-HUB (programme 2014–2020). Our gratitude extends also to Jürgen Finsterbusch for support on the MR side, to the University of Minnesota Centre for Magnetic Resonance Research for providing the image reconstruction algorithm for the simultaneous multislice acquisitions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.674719/full#supplementary-material

REFERENCES

André, E. D., Grinberg, F., Farrher, E. I, Maximov, I., Shah, N. J., Meyer, C., et al. (2014). Influence of noise correction on intra- and inter-subject variability of quantitative metrics in diffusion kurtosis imaging. *PLoS One* 9:e94531. doi: 10.1371/journal.pone.0094531

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Berman, S., West, K. L., Does, M. D., Yeatman, J. D., and Mezer, A. A. (2018). Evaluating g-ratio weighted changes in the corpus callosum as a function of age and sex. *Neuroimage* 182, 304–313. doi: 10.1016/j.neuroimage.2017.06.076

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310. doi: 10.1128/AAC.00483-18

Callaghan, M. F., Helms, G., Lutti, A., Mohammadi, S., and Weiskopf, N. (2015a). A general linear relaxometry model of R1 using imaging data. *Magn. Reson. Med.* 73, 1309–1314. doi: 10.1002/mrm.25210

Callaghan, M. F., Josephs, O., Herbst, M., Zaitsev, M., Todd, N., and Weiskopf, N. (2015b). An evaluation of prospective motion correction (PMC) for high resolution quantitative MRI. *Front. Neurosci.* 9:97. doi: 10.3389/fnins.2015.00097

Callaghan, M. F., Lutti, A., Ashburner, J., Balteau, E., Corbin, N., Draganski, B., et al. (2019). Example dataset for the hMRI toolbox. *Data Brief* 25:104132. doi: 10.1016/j.dib.2019.104132

Campbell, J. S. W., Leppert, I. R., Boudreau, M., Narayanan, S., Duval, T., Cohen-Adad, J., et al. (2018). Promise and pitfalls of g-ratio estimation with MRI. *Neuroimage* 182, 80–96. doi: 10.1016/j.neuroimage.2017.08.038

Cercignani, M., Giulietti, G., Dowell, N. G., Gabel, M., Broad, R., Leigh, P. N., et al. (2017). Characterizing axonal myelination within the healthy population: a tract-by-tract mapping of effects of age and gender on the fiber g-ratio. *Neurobiol. Aging* 49, 109–118. doi: 10.1016/j.neurobiolaging.2016.09.016

Ellerbrock, I., and Mohammadi, S. (2018). Four in vivo g -ratio-weighted imaging methods: comparability and repeatability at the group level. *Hum. Brain Mapp.* 39, 24–41. doi: 10.1002/hbm.23858

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., et al. (2010). Multiplexed echo planar imaging for sub-second whole brain FMRI and Fast Diffusion Imaging. Edited by Pedro Antonio Valdes-Sosa. *PLoS One* 5:e15710. doi: 10.1371/journal.pone.0015710

Gloor, M., Scheffler, K., and Bieri, O. (2008). Quantitative magnetization transfer imaging using balanced SSFP. *Magn. Reson. Med.* 60, 691–700. doi: 10.1002/mrm.21705

Graf von Keyserlingk, D., and Schramm, U. (1984). Diameter of axons and thickness of myelin sheaths of the pyramidal tract fibres in the adult human medullary pyramid. *Anat. Anz.* 157, 97–111. doi: 10.1097/PSY.0000000000000183

Helms, G. (2015). "Correction for residual effects of B1+ inhomogeniety on MT saturation in FLASH-based multi-parameter mapping of the brain," in *Proceedings of the International Society for Magnetic Resonance in Medicine*

Annual Meeting 2015, (Toronto: International Society for Magnetic Resonance in Medicine), 3360.

Helms, G., Dathe, H., Kallenberg, K., and Dechent, P. (2008). High–resolution maps of magnetization transfer with inherent correction for RF inhomogeneity and T 1 relaxation obtained from 3D FLASH MRI. *Magn. Reson. Med.* 60, 1396–1407. doi: 10.1002/mrm.21732

Helms, G., Weiskopf, N., and Lutti, A. (2021). Correction of FLASH-based MT saturation in human brain for residual bias of B1-inhomogeneity at 3T. *arXiv* [Preprint]. arXiv:210414878 [physics].

Hildebrand, C., and Hahn, R. (1978). Relation between myelin sheath thickness and axon size in spinal cord white matter of some vertebrate species. *J. Neurol. Sci.* 38, 421–434. doi: 10.1016/0022-510X(78)90147-8

Hori, M., Hagiwara, A., Fukunaga, I., Ueda, R., Kamiya, K., Suzuki, Y., et al. (2018). Application of quantitative microstructural MR imaging with atlas-based analysis for the spinal cord in cervical spondylotic myelopathy. *Sci. Rep.* 8, 1–8. doi: 10.1038/s41598-018-23527-8

Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D. S., et al. (2008). Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *Neuroimage* 39, 336–347. doi: 10.1016/j.neuroimage.2007.07.053

Jung, W., Lee, J., Shin, H. G., Nam, Y., Zhang, H., Oh, S. H., et al. (2017). Whole brain g-ratio mapping using myelin water imaging (MWI) and neurite orientation dispersion and density imaging (NODDI). *Neuroimage* 19, 109–118. doi: 10.1016/j.neuroimage.2017.09.053

Kamagata, K., Zalesky, A., Yokoyama, K., Andica, C., Hagiwara, A., Shimoji, K., et al. (2019). MR g-ratio-weighted connectome analysis in patients with multiple sclerosis. *Sci. Rep.* 9:13522. doi: 10.1038/s41598-019-50025-2

LaMantia, A.-S., and Rakic, P. (1990). Cytological and quantitative characteristics of four cerebral commissures in the rhesus monkey. *J. Comp. Neurol.* 291, 520–537. doi: 10.1002/cne.902910404

Lehmenkühler, A., Syková, E., Svoboda, J., Zilles, K., and Nicholson, C. (1993). Extracellular space parameters in the rat neocortex and subcortical white matter during postnatal development determined by diffusion analysis. *Neuroscience* 55, 339–351. doi: 10.1016/0306-4522(93)90503-8

Leutritz, T., Seif, M., Helms, G., Samson, R. S., Curt, A., Freund, P., et al. (2020). Multiparameter mapping of relaxation (R1, R2*), proton density and magnetization transfer saturation at 3 T: a multicenter dual-vendor reproducibility and repeatability study. *Hum. Brain Mapp.* 41, 4232–4247. doi: 10.1002/hbm.25122

Liewald, D., Miller, R., Logothetis, N., Wagner, H.-J., and Schüz, A. (2014). Distribution of axon diameters in cortical white matter: an electron-microscopic study on three human brains and a macaque. *Biol. Cybern.* 108, 541–557. doi: 10.1007/s00422-014-0626-2

Lutti, A., Hutton, C., Finsterbusch, J., Helms, G., and Weiskopf, N. (2010). Optimization and validation of methods for mapping of the radiofrequency transmit field at 3T. *Magn. Reson. Med.* 238, 229–238. doi: 10.1002/mrm.22421

Macdonald, J., and Ruthotto, L. (2018). Improved susceptibility artifact correction of echo-planar MRI using the alternating direction method of multipliers. *J. Math. Imaging Vis.* 60, 268–282. doi: 10.1007/s10851-017-0757-x

Mancini, M., Giulietti, G., Dowell, N., Spanò, B., Harrison, N., Bozzali, M., et al. (2017). Introducing axonal myelination in connectomics: a preliminary analysis of g-ratio distribution in healthy subjects. *Neuroimage* 182, 351–359. doi: 10.1016/j.neuroimage.2017.09.018

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., et al. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain FMRI. *Magn. Reson. Med.* 63, 1144–1153. 10.1002/mrm.22361

Mohammadi, S., and Callaghan, M. F. (2020). Towards in vivo g-ratio mapping using MRI: unifying myelin and diffusion imaging. *J. Neurosci. Methods* 348:108990. doi: 10.1016/j.jneumeth.2020.108990

Mohammadi, S., Carey, D., Dick, F., Diedrichsen, J., Sereno, M. I., Reisert, M., et al. (2015). Whole-brain in-vivo measurements of the axonal g-ratio in a group of 37 healthy volunteers. *Front. Neurosci.* 9:441. doi: 10.3389/fnins.2015.00441

Mohammadi, S., Möller, H. E., Kugel, H., Müller, D. K., and Deppe, M. (2010). Correcting eddy current and motion effects by affine whole-brain registrations: evaluation of three-dimensional distortions and comparison with slicewise correction. *Magn. Reson. Med.* 64, 1047–1056. doi: 10.1002/mrm.22501

Mohammadi, S., Tabelow, K., Ruthotto, L., Feiweier, T., Polzehl, J., and Weiskopf, N. (2014). High-resolution diffusion kurtosis imaging at 3t enabled by advanced post-processing. *Front. Neurosci.* 8:427. doi: 10.3389/fnins.2014.00427

Nicholson, C., and Hrabìtová, S. (2017). Brain extracellular space: the final frontier of neuroscience. *Biophys. J.* 113, 2133–2142. doi: 10.1016/j.bpj.2017.06.052

Olivares, R., Montiel, J., and Aboitiz, F. (2001). Species differences and similarities in the fine structure of the mammalian corpus callosum. *Brain Behav. Evol.* 57, 98–105. doi: 10.1159/000047229

Reese, T. G., Heid, O., Weisskoff, R. M., and Wedeen, V. J. (2003). Reduction of eddy-current-induced distortion in diffusion MRI using a twice-refocused spin echo. *Magn. Reson. Med.* 49, 177–182. doi: 10.1002/mrm.10308

Ruthotto, L., Kugel, H., Olesch, J., Fischer, B., Modersitzki, J., Burger, M., et al. (2012). Diffeomorphic susceptibility artifact correction of diffusion-weighted magnetic resonance images. *Phys. Med. Biol.* 57, 5715–5731. doi: 10.1088/0031-9155/57/18/5715

Ruthotto, L., Mohammadi, S., Heck, C., Modersitzki, J., and Weiskopf, N. (2013). "Hyperelastic susceptibility artifact correction of DTI in SPM," in *Bildverarbeitung Für Die Medizin 2013*, eds H. P. Meinzer, T. Deserno, H. Handels, and T. Tolxdorff (Berlin: Springer), 344–349. doi: 10.1007/978-3-642-36480-8_60

Sereno, M. I., Lutti, A., Weiskopf, N., and Dick, F. (2013). Mapping the human cortical surface by combining quantitative T1 with retinotopy. *Cereb. Cortex* 23, 2261–2268. doi: 10.1093/cercor/bhs213

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald, L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magn. Reson. Med.* 67, 1210–1224. doi: 10.1002/mrm.23097

Stikov, N., Campbell, J. S., Stroh, T., Lavelée, M., Frey, S., Novek, J., et al. (2015). In vivo histology of the myelin g-ratio with magnetic resonance imaging. *Neuroimage* 118, 397–405. doi: 10.1016/j.neuroimage.2015.05.023

Stikov, N., Perry, L. M., Mezer, A., Rykhlevskaia, E., Wandell, B. A., Pauly, J. M., et al. (2011). Bound pool fractions complement diffusion measures to describe white matter micro and macrostructure. *Neuroimage* 54, 1112–1121. doi: 10.1016/j.neuroimage.2010.08.068

Swadlow, H. A., Waxman, S. G., and Geschwind, N. (1980). Small-diameter nonmyelinated axons in the primate corpus callosum. *Arch. Neurol.* 37, 114–115. doi: 10.1001/archneur.1980.00500510072016

Tabelow, K., Balteau, E., Ashburner, J., Callaghan, M. F., Draganski, B., Helms, G., et al. (2019). hMRI – A toolbox for quantitative MRI in neuroscience and clinical research. *Neuroimage* 194, 191–210.

Tønnesen, J., Inavalli, V. V. G. K., and Nägerl, U. V. (2018). Super-resolution imaging of the extracellular space in living brain tissue. *Cell* 172, 1108–1121. doi: 10.1016/j.cell.2018.02.007

Wang, S. S., Shultz, J. R., Burish, M. J., Harrison, K. H., Hof, P. R., Towns, L. C., et al. (2008). Functional trade-offs in white matter axonal scaling. *J. Neurosci.* 28, 4047–4056. doi: 10.1523/JNEUROSCI.5559-05.2008

Weiskopf, N., Lutti, A., Helms, G., Novak, M., Ashburner, J., and Hutton, C. (2011). Unified segmentation based correction of R1 brain maps for RF transmit field inhomogeneities (UNICORT). *Neuroimage* 54, 2116–2124. doi: 10.1016/j.neuroimage.2010.10.023

Weiskopf, N., Suckling, J., Williams, G., Correia, M. M., Inkster, B., Tait, R., et al. (2013). Quantitative multi-parameter mapping of R1, PD(*), MT, and R2(*) at 3T: a multi-center validation. *Front. Neurosci.* 7:95. doi: 10.3389/fnins.2013.00095

West, K. L., Kelm, N. D., Carson, R. P., and Does, M. D. (2016). A revised model for estimating g-ratio from MRI. *Neuroimage* 125, 1155–1158. doi: 10.1016/j.neuroimage.2015.08.017

Xu, J., Moeller, S., Auerbach, E. J., Strupp, J., Smith, S. M., Feinberg, D. A., et al. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3T. *Neuroimage* 83, 991–1001. doi: 10.1016/j.neuroimage.2013.07.055

Yu, F., Fan, Q., Tian, Q., Ngamsombat, C., Machado, N., Bireley, J. D., et al. (2019). Imaging g-ratio in multiple sclerosis using high-gradient diffusion mri and macromolecular tissue volume. *AJNR Am. J. Neuroradiol.* 40, 1871–1877. doi: 10.3174/ajnr.A6283

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage* 61, 1000–1016. doi: 10.1016/j.neuroimage.2012.03.072

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 31 August 2021
doi: 10.3389/fnins.2021.708196

A Contrast Augmentation Approach to Improve Multi-Scanner Generalization in MRI

Maria Ines Meyer[1,2]*, *Ezequiel de la Rosa*[2,3], *Nuno Pedrosa de Barros*[2], *Roberto Paolella*[2,4], *Koen Van Leemput*[1,5] *and Diana M. Sima*[2]

[1] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark, [2] Icometrix, Leuven, Belgium, [3] Department of Computer Science, Technical University of Munich, Munich, Germany, [4] Imec Vision Lab, University of Antwerp, Antwerp, Belgium, [5] Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States

Most data-driven methods are very susceptible to data variability. This problem is particularly apparent when applying Deep Learning (DL) to brain Magnetic Resonance Imaging (MRI), where intensities and contrasts vary due to acquisition protocol, scanner- and center-specific factors. Most publicly available brain MRI datasets originate from the same center and are homogeneous in terms of scanner and used protocol. As such, devising robust methods that generalize to multi-scanner and multi-center data is crucial for transferring these techniques into clinical practice. We propose a novel data augmentation approach based on Gaussian Mixture Models (GMM-DA) with the goal of increasing the variability of a given dataset in terms of intensities and contrasts. The approach allows to augment the training dataset such that the variability in the training set compares to what is seen in real world clinical data, while preserving anatomical information. We compare the performance of a state-of-the-art U-Net model trained for segmenting brain structures with and without the addition of GMM-DA. The models are trained and evaluated on single- and multi-scanner datasets. Additionally, we verify the consistency of test-retest results on same-patient images (same and different scanners). Finally, we investigate how the presence of bias field influences the performance of a model trained with GMM-DA. We found that the addition of the GMM-DA improves the generalization capability of the DL model to other scanners not present in the training data, even when the train set is already multi-scanner. Besides, the consistency between same-patient segmentation predictions is improved, both for same-scanner and different-scanner repetitions. We conclude that GMM-DA could increase the transferability of DL models into clinical scenarios.

Keywords: multi-scanner, magnetic resonance imaging, segmentation, data augmentation, gaussian mixture models

OPEN ACCESS

Edited by:
Nico Sollmann, University of California, San Francisco, United States

Reviewed by:
Matthan Caan, University of Amsterdam, Netherlands
Blaise Frederick, Harvard Medical School, United States

***Correspondence:**
Maria Ines Meyer
ines.meyer@icometrix.com; mifm@dtu.dk

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *11 May 2021*
Accepted: *27 July 2021*
Published: *31 August 2021*

Citation:
Meyer MI, de la Rosa E, Pedrosa de Barros N, Paolella R, Van Leemput K and Sima DM (2021) A Contrast Augmentation Approach to Improve Multi-Scanner Generalization in MRI. Front. Neurosci. 15:708196. doi: 10.3389/fnins.2021.708196

1. INTRODUCTION

The segmentation of different brain structures from Magnetic Resonance Imaging (MRI) is an important problem in the field of neuroimaging. Obtaining precise and consistent delineations is crucial in the diagnosis, follow-up and treatment of neurological disorders. Important examples are the monitoring of the progression of Multiple Sclerosis (MS) or dementia, both connected

to accentuated neurodegeneration (Giorgio and De Stefano, 2013). In recent years, convolutional neural networks (CNN) have become an increasingly popular segmentation approach, but the fact that these models are extremely sensitive to data variability has hindered their large scale adoption in clinical and research settings. Specifically, these algorithms remain sensitive to factors such as hardware and acquisition settings, which can be especially problematic when integrating data from different cohorts (Mårtensson et al., 2020). For these models to generalize to data collected using new or unseen scanners, large multi-center and multi-scanner datasets are necessary at the training stage. Nevertheless, collecting such data is not trivial and most available datasets are homogeneous in terms of scanner types and acquisition protocols.

1.1. Related Work

The above mentioned problem is often termed as the *scanner bias* problem. A popular way to deal with it in large clinical trials is through approaches based on statistical *harmonization*. In most cases the focus is on removing the scanner bias from the volumetric measurements based on scanner- or center-information (Fortin et al., 2018; Garcia-Dias et al., 2020). At the image level, it is common to use the standardization of the MRI intensity scale to reduce scanner sensitivity (Wang et al., 1998; Nyúl and Udupa, 1999; Shinohara et al., 2014), which has been previously shown to improve the outcome of computer vision tasks like segmentation (Zhuge and Udupa, 2009) and registration (Bagci et al., 2010). Recently, some works have attempted to use Deep Learning (DL) methods to modify the analyzed images such that they appear to have been acquired under similar settings (Dewey et al., 2019; Zhao et al., 2019b). However, harmonization methods have the undesirable property that the results will always be bound by the least informative scanner in the dataset, as shown in Moyer and Golland (2021), while standardization methods are not able to remove residual across-subject variability (Shinohara et al., 2014; Fortin et al., 2016; Wrobel et al., 2020). Additionally, many of these approaches require retraining and updating of the models when including new data from unseen scanners or centers.

In order to avoid these unwanted effects, it is interesting to tackle the problem from a generalization perspective, by improving the performance and reproducibility of the methods of interest (often segmentation of brain tissues or lesions). When considering DL methods in particular, a common approach is to increase the variability in the data by applying well designed data augmentation (DA). The idea behind DA is simple: by applying transformations to the labeled data it is possible to artificially increase the training set, which implicitly regularizes the trained CNN. The most common DA strategies explore transformations of the original data, mostly based on the application of operations such as elastic distortions (Simard et al., 2003), linear geometric transformations such as translations and rotations, color transformations (mostly by altering the intensities of the RGB channels in 2D images) (Krizhevsky et al., 2012) or noise injection (Sietsma and Dow, 1991).

In the medical imaging field, DA is especially important since annotated datasets are typically small. Although simple transformations such as the ones described above can alleviate overfitting and improve performance on the test sets in medical applications (Milletari et al., 2016), they do not take into account the high variability in terms of contrast found in MRI. Some works have attempted to overcome this limitation by generating completely synthetic images using generative adversarial networks, as is the case in Shin et al. (2018). Nonetheless, there is still a long way to go until these images can be used effectively. Other more promising approaches start from existing images and alter them in such a way that new sequences or contrasts are simulated. One relevant example is described in Jog et al. (2019), where a CNN-based algorithm resilient to variations in the input acquisition is presented. To achieve this, approximate forward models of different MRI pulse sequences are built. This way, synthetic versions of the training images are generated such that they appear to have been acquired using different sequences. The method has the disadvantage that it is complex, slow and it requires nuclear magnetic resonance parameter maps of the training images, which are often unavailable. Zhao et al. (2019a) proposed to learn a model of transformations from an atlas to images in a dataset and to use this model along with a single labeled example to synthesize additional labeled examples with variable appearance and spacial deformations. More recently, Billot et al. (2020) presented a contrast-agnostic brain segmentation method, again based on generating synthetic images. The method uses only a segmentation map to generate new images with varying, sometimes even unrealistic, contrasts. The generated images have random appearance, deformation, noise, and bias field. With this type of extreme augmentation, it is possible to obtain a final model that is not biased toward any specific MRI contrast and that achieves good performance on unpreprocessed brain scans of any contrast. Although this method is very promising, by design it is limited to segmentation applications and nuanced variations in the individual images are lost.

1.2. Our Contribution

In the present work we propose a novel intensity-based DA strategy with the main goal of reducing the scanner bias of models trained on data with low protocol-, scanner- or center-variability. Although scanner factors cause variations to other image characteristics (e.g., noise, artifacts, geometric distortions), we have previously found a clear relationship between tissue contrast and volume measurements (Meyer et al., 2019). As such, we hypothesize that augmenting the tissue intensity variability will have a positive effect in the model generalization to new, unseen scanners or center-specific acquisition configurations. The method is based on the Gaussian Mixture Model (GMM) framework: we estimate the individual tissue components of an MRI image and randomly modify them, while preserving structural information. As a result the contrast between different tissues varies, in a similar way to what happens when different scanners or sequences are used during acquisition. We validate the approach in the task of brain structure segmentation. Unlike currently existing methods, the proposed approach does not depend on any existent segmentations or parameter maps; it is

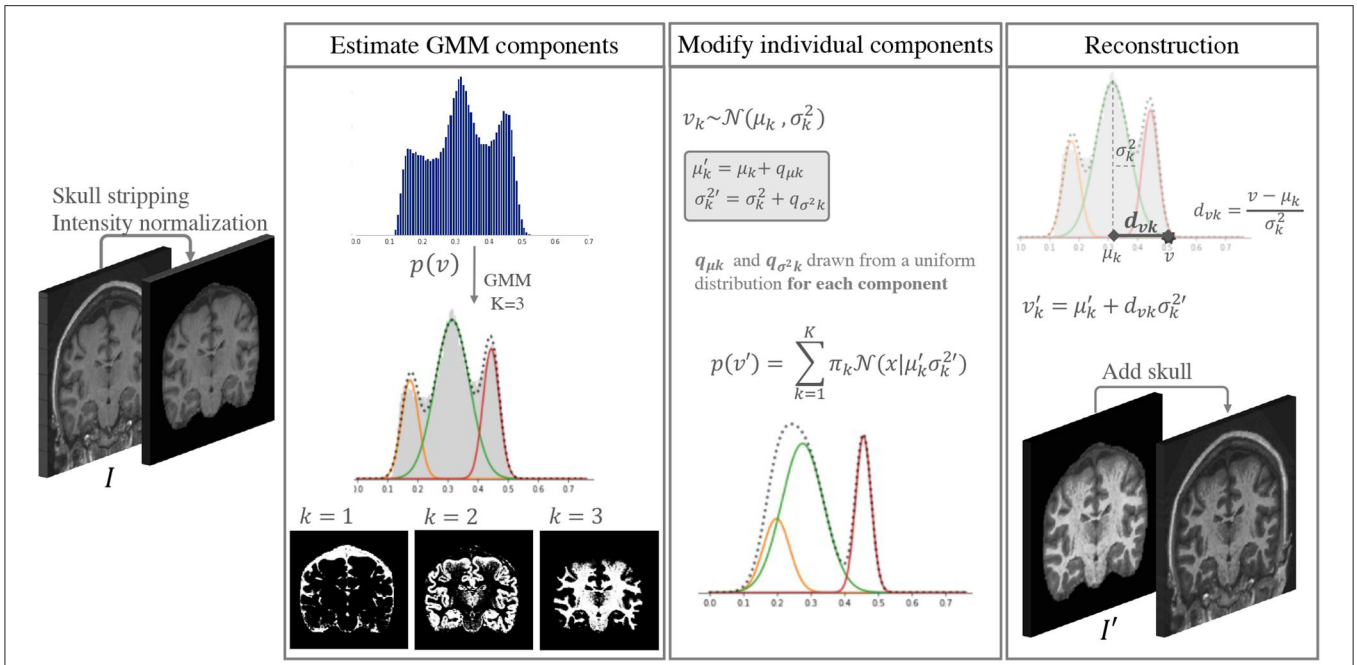


FIGURE 1 | Diagram of the main steps in the proposed DA method. This augmentation is performed only while training the segmentation network.

simple and fast; it can be used on-the-fly during training; and it is not necessarily limited to segmentation applications.

This work extends the preliminary research presented in Meyer et al. (2021). We previously compared the performance of the same CNN-based model trained under three different settings: (i) single-scanner data, (ii) single-scanner data with the addition of our DA method, and (iii) multi-scanner data. We now additionally investigate the effect of adding the DA method to multi-scanner data and evaluate the reproducibility and consistency of the models on a *test-retest* dataset containing same-patient repetitions in the same and different scanners. Finally, we investigate the effect of the presence of bias field on the training images. Overall we observe a clear improvement in generalization to unseen scanner types when adding the proposed method to the training pipeline, not only when the original training dataset is homogeneous, but also in the case when a large, heterogeneous dataset is used as training set.

2. GAUSSIAN MIXTURE MODEL-BASED INTENSITY TRANSFORMATION

The idea behind the proposed approach is to increase the intensity and contrast variability of images in datasets with low scanner and center acquisition diversity, such that it becomes representative of what is found in large multi-scanner and multi-center cohorts. This DA method is applied during the training phase of a DL network of choice, and is not necessary at inference. **Figure 1** shows a depiction of the method. An implementation is available at https://github.com/icometrix/gmm-augmentation.

2.1. The Gaussian Mixture Model Framework

It is well documented that in a skull-stripped T1w brain MRI without contrast injection, characteristic peaks in the histogram correspond to different tissues, i.e., CSF has the lowest intensity, followed by GM and WM. This has been explored by several segmentation methods based on Gaussian Mixture Models (GMM) (Van Leemput et al., 1999; Ashburner and Friston, 2005). GMM is a type of probabilistic model that assumes that data can be modeled as a superposition of K Gaussians. Within this framework, if we have a set of observations $\{v_1, \ldots v_N\}$, corresponding to the intensities v of each voxel $n \in N$ in an image I, we can model each observation in the data using a mixture of Gaussians, such that:

$$p(v_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(v_n|\mu_k, \sigma_k^2). \quad (1)$$

Each $\mathcal{N}(\mu_k, \sigma_k^2)$ is a *component* of the mixture, with its own mean μ_k and variance σ_k^2, and π_k are the mixing coefficients. For simplicity we hide the subscript n when referring to the intensity of a given voxel: v_n is represented as v from here on.

We start by selecting $K = 3$ Gaussian components for the GMM, where each component roughly corresponds to the CSF, GM, and WM classes. The parameters are initialized and updated iteratively using the Expectation Maximization (EM) (Dempster et al., 1977) algorithm implemented in the scikit-learn package for Python (Pedregosa et al., 2011) with default parameters.

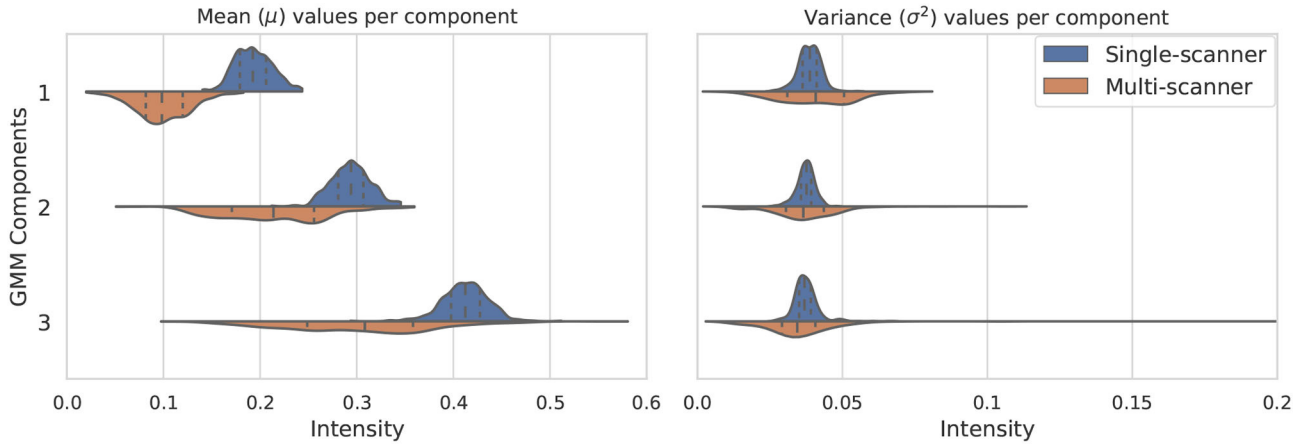


FIGURE 2 | Variation of the three considered GMM components in terms of mean μ_k (left) and variance σ_k^2 (right). The components vary much more in a multi-scanner dataset than in a single-scanner setting.

Once we estimate the parameters for each component k, we can use Bayes' rule to compute the probability of each class label C, such that:

$$p(C = k|v) = \frac{\pi_k \mathcal{N}(v|\mu_k, \sigma_k^2)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(v|\mu_{k'}, \sigma_{k'}^2)}. \quad (2)$$

2.2. Altering the Components of the GMM

If we modify the individual components of a 3-component GMM we can modify images in the training data by changing their GMM probability distributions while preserving the inherent image characteristics. We can create a new intensity distribution for each of the tissues by generating new parameters $\mu_k \rightarrow \mu_k'$ and $\sigma_k^2 \rightarrow \sigma_k^{2\prime}$ for each of the components in an individual skull stripped image. To do this we:

a) sample individual variation terms q_{μ_k} and $q_{\sigma_k^2}$ for each component from a uniform distribution,
b) add these values to the original parameters, such that $\mu_k' = \mu_k + q_{\mu_k}$ and $\sigma_k^{2\prime} = \sigma_k^2 + q_{\sigma_k^2}$.

To define the range of the uniform distributions we use to sample the variation terms q_{μ_k} and $q_{\sigma_k^2}$, we start by estimating the range of typical variation for each component from a large multi-scanner collection of patient data (dataset C in section 3). To do this, all images are first skull stripped, intensities are clipped at percentiles 1 and 99 to remove extreme values, and normalized to the range $[0, 1]$. Then we fit a 3-component GMM to each image in the dataset using the same procedure as described above. We extract the mean μ_k and variance σ_k^2 values of each component. We then use the standard deviation ($s(\cdot)$) of the estimated parameters to define the range of variability we allow. q_{μ_k} and $q_{\sigma_k^2}$ are sampled for each component from the uniform distributions $U(-s(\mu_k), s(\mu_k))$ and $U(-s(\sigma_k^2), s(\sigma_k^2))$, respectively.

The distribution of the estimated parameters μ_k and σ_k^2 is depicted in **Figure 2**. In this figure we illustrate how the variability of the estimated parameters in a multi-scanner and multi-center setting is larger than that of a homogeneous dataset (same center, same scanner, same acquisition protocol) (dataset A in section 3). Besides differences in hardware, acquisitions in different centers tend to not be perfectly harmonized, which causes variations in contrast of the images. This is one of the many factors that contribute to the increased variability of the estimated parameters, and can be addressed by the proposed approach. For the multi-center data, mean and variance values for the 3 components have approximate standard deviations of $s(\mu) = \{3, 6, 8\} \times 10^{-2}$ and $s(\sigma^2) = \{1, 1, 3\} \times 10^{-3}$, respectively.

The choice of a uniform distribution for sampling the new variation terms implies that any random combination of tissue intensities can be generated. We could restrict this to more probable distributions by selecting a normal distribution. However, since exposing networks to extreme but anatomically plausible augmentation can be beneficial for learning (Billot et al., 2020), we decided to allow the possibility for some unrealistic combinations to arise.

2.3. Reconstruction

Once the new parameters have been defined, we could think that a logical next step would be to generate a new histogram of intensities by mixing the new Gaussian distributions and using histogram matching (Wang et al., 1998) techniques to generate a new image I'. However, doing this would not guarantee that structural information is preserved (e.g., two components could overlap or even shift order, and voxels from one tissue would be wrongly assigned to another class). To avoid this we describe the intensity v of some voxel $n \in N$ in terms of the distance from the mean of the component measured with the Mahalanobis distance $d_{vk} = (v - \mu_k)/\sigma_k$. This implies that if we know the values of μ_k and σ_k^2 we can find the updated value of $v \rightarrow v'$ for each component k by preserving the distance d_{vk}:

$$v_k' = \mu_k' + d_{vk}\sigma_k'. \quad (3)$$

Finally, we can compute the new intensity v' for a voxel n by leveraging each component by the initial probability

that this voxel belonged to a certain class $p(C = k|v)$, such that

$$v' = \sum_{k=1}^{N} p(C = k|v)v'_k. \quad (4)$$

This guarantees that the voxels that have a high probability of belonging to a certain class will represent the same class, while allowing for nuanced variations at the borders between different tissues.

3. DATASETS AND EXPERIMENTAL SETUP

From here onwards, the method will be referred to as *GMM-DA*, for simplicity. In order to validate the GMM-DA method, we investigate the added value of the described method on the task of brain structure segmentation using a well described type of CNN architecture. We train the same network on two different datasets: a collection of single-scanner data from healthy subjects, and a multi-scanner and multi-center collection of patient data. We compare the performance of the models trained with and without the addition of the GMM-DA strategy. The different models are evaluated on manual segmentations and on test-retest data. The available datasets and the different experiments are described in the following sections.

3.1. Available Datasets

A) OASIS

Contains T1w MRI scans from 416 subjects (age: [18, 96] years) (the OASIS-1 cohort) (Marcus et al., 2007). Only 280 of the 316 healthy subjects were considered (see dataset B). The data was randomly split into train/validation/test sets [$n = 179(64\%)/45(16\%)/56(20\%)$]. Although the data is heterogeneous from a population point of view, it is extremely homogeneous in terms of protocol and scanner. All images were acquired on a 1.5T Siemens Vision scanner, using the MP-RAGE sequence with constant repetition time (TR) and echo time (TE) (TR: 9.7 ms; TE: 4.0 ms). Slice thickness is also constant (1.25 mm).

B) MICCAI 2012

Contains 35 T1w scans from healthy subjects. The original MRI scans are from OASIS, but this dataset contains manual labels of brain structures. These data were provided for use in the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling (Landman and Warfield, 2012). All the images in this dataset were removed from OASIS prior to splitting the data into the different training and test sets, to avoid overlap. We exclude 5 scans from repeated subjects and use the remaining 30 for evaluating the methods on the manual labels.

C) MS Dataset

This is a collection of multi-center T1w MRI scans from 421 individual Multiple Sclerosis (MS) patients. It contains a lot of variability both at the population level and in terms of scanner- and center- or acquisition-specific factors, i.e., age ([16, 81] years), sex (M/F $\sim$ 33%/67%), slice thickness in T1 ([0.4, 1.5] mm), magnetic field strength (1.5T/3T $\sim$ 43%/57%), scanner manufacturer (Philips, GE, Siemens and Hitachi), scanner model (29 devices) and acquisition sequence (TR: [4.9, 5000] ms; TE: [1.9, 8.0] ms). This dataset, which we term *heterogeneous*, was used to estimate the range of typical variation of the GMM components for the different tissues, as described in section 2.2. Additionally, we used this data to generate an *independent test set*, containing 92 images from 10 different scanner models. For an additional experiment we pooled a train/validation set of 251/44 images, ensuring that any scanner models present in the pre-selected test set or in OASIS were not included.

D) Test-Retest Dataset

Contains T1w MRI scans from 10 MS patients. Each patient was scanned twice (with re-positioning) on three different 3T scanner types with different acquisition sequences: (i) Philips Achieva: 3D T1-weighted FSPGR sequence (TR 4.93 ms); (ii) Siemens Skyra: 3D T1-weighted MP-RAGE sequence (TR 2300 ms, TE 2.29 ms); (iii) GE Discovery MR450w: 3D T1-weighted FSPGR sequence (TR 7.32 ms, TE 3.14 ms). Further details regarding this data can be found in Jain et al. (2015). This dataset allows the models to be tested for consistency, both in an *intra-scanner* setting as well as in an *inter-scanner* setting.

3.2. Data Pre-processing

All images were normalized using a modified z-score function robust against outliers, where the median of the distribution was preferred over of the mean, and the standard deviation of the distribution was computed within percentiles 10 and 90. Additionally, images were bias-field corrected using the N4 inhomogeneity correction algorithm as implemented in the Advanced Normalization Tools (ANTs) toolkit (Tustison et al., 2010) and linearly registered to MNI space using the tools implemented in NiftyReg (Ourselin et al., 2001).

3.3. Experimental Setup

We trained a CNN to segment White Matter (WM), Gray Matter (GM), Cerebro-Spinal Fluid (CSF), Lateral Ventricles (LV), Thalamus (Tha), Hippocampus (HC), Caudate Nucleus (CdN), Putamen (Pu) and Globus Palidus (GP). Due to scarcity of manual delineations, we train and evaluate the CNN models using brain substructure delineations obtained with ico**brain** (Jain et al., 2015; Struyfs et al., 2020), a clinically available and FDA-approved Software.

3.4. Model Architecture

For the segmentation task we use a 3D UNet architecture (Çiçek et al., 2016) with a few adaptations, namely:

- Weight normalization layers (Salimans and Kingma, 2016) are added after each convolutional operation instead of batch normalization;
- LeakyReLU (Maas et al., 2013) is used as the main activation function;
- Strided convolutions are used instead of max pooling.

The models are trained using a combination of the soft-dice loss ($\mathcal{L}_{Dice}$) and the weighted categorical cross-entropy loss ($\mathcal{L}_{wCE}$), as

TABLE 1 | Summary of the trained models.

Model types	Training/Testing datasets		Testing datasets	
	OASIS	MS dataset	MICCAI 2012	Test-retest
CNN_{OASIS}	T, E	E	E	E
$CNN_{OASIS\text{-}DA}$	T, E	E	E	E
CNN_{MS}	E	T, E	E	E
$CNN_{MS\text{-}DA}$	E	T, E	E	E
$CNN_{MS\text{-}BF}$	-	T, E	-	-
$CNN_{MS\text{-}BF\text{-}DA}$	-	T, E	-	-

T, trained; E, evaluated.

suggested in Isensee et al. (2021):

$$\mathcal{L}_{total} = \mathcal{L}_{wCE} + \mathcal{L}_{Dice}. \tag{5}$$

Considering that $y_n \in \{0, 1\}$ is the one-hot-encoded label of the n^{th} voxel in the model's input and $\hat{y}_n \in [0, 1]$ is the prediction output of the model for the same voxel, the soft-Dice loss is an extension to K classes of the popular Dice loss presented, as presented in Sudre et al. (2017):

$$\mathcal{L}_{Dice} = 1 - 2\frac{\sum_{k=1}^{K}\sum_{n}\hat{y}_{nk}y_{nk}}{\sum_{k=1}^{K}\sum_{n}\hat{y}_{nk} + y_{nk}}. \tag{6}$$

To deal with the accentuated class imbalance of this problem we use the weighted categorical cross-entropy loss similarly to what was described in Ronneberger et al. (2015). This loss function can be expressed as:

$$\mathcal{L}_{wCE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} w_{nk}y_{nk}\log\hat{y}_{nk}, \tag{7}$$

where w_{nk} is the weighting factor for the n-th voxel belonging to the k-class in the training set. These weights allow to compensate the scarcity of voxels from some of the classes.

The network takes as input patches of size $128 \times 128 \times 128$ and outputs probability maps of size $88 \times 88 \times 88$. Kernel size is $3 \times 3 \times 3$ and initial number of filters 16 (raised to the power of 2 at each layer in the encoder path). The model is implemented using Tensorflow 2.0 and trained until convergence using mini-batch stochastic gradient descent (Adam optimizer) with initial learning rate $\lambda = 0.001$ on a machine equipped with a Tesla K80 Nvidia GPU (12 GB dedicated).

3.5. Experiments

To validate the approach we compare the performance of models trained with and without the addition of the GMM-DA strategy. First, we evaluate how a model trained on single scanner data generalizes to an unseen multi-scanner dataset (train on the OASIS training set, and evaluate on the OASIS test set and the MS dataset test set). This is the key experiment in the results, since we are particularly interested in evaluating the increase in generalizability of the CNN to multi-scanner and multi-center data after adding the augmentation step. Although we acknowledge the presence of white matter lesions in the images from the MS dataset, we decide not to deal with them explicitly in this context. Secondly, in order to evaluate how the same network performs on unseen scanners and centers when trained on heterogeneous data, we train the same models on the MS dataset described in section 3. We additionally investigate if the addition of GMM-DA in this setting is still beneficial. We proceed to compare these four approaches on manual labels and on the test-retest dataset. Finally, we evaluate how the presence of bias field (BF) on the training images impacts the performance of the GMM-DA. To this end, we train the same models on the MS dataset images, this time without the bias field correction step.

We train and evaluate a total of six models. The models are named according to the architecture (CNN), training data (OASIS or MS), presence of bias field (BF) on the training images and addition of the data augmentation (DA) step. As such, a model trained on the MS dataset, on data with bias field and to which GMM-DA was applied is termed $CNN_{MS\text{-}BF\text{-}DA}$. The investigated models and a description of the data where they were trained (T) or evaluated (E) are summarized in **Table 1**.

3.6. Performance Metrics

Dice scores (*DC*), sensitivity (*Se*) and precision (*Pr*) are reported (complete *Se* and *Pr* results are given in the **Supplementary Material**). *DC* values are compared using Wilcoxon paired rank-sum and Levene tests to evaluate the null hypotheses H_0 that the results from the different models have equal median and variance values, respectively. These tests were selected given the presence of outliers and deviations from normality in the distributions (see **Figure 3**). When evaluating the reproducibility of the methods, absolute volume differences between acquisitions from the same patient are reported. Results are summarized in terms of median ($P50$) and percentile 10 ($P10$) or 90 ($P90$), where relevant.

4. EXPERIMENTS AND RESULTS

4.1. GMM Augmentation of a Homogeneous Dataset

To evaluate the influence of the addition of GMM augmentation when training on a homogeneous dataset (OASIS), we test CNN_{OASIS} and $CNN_{OASIS\text{-}DA}$ on the two cross sectional datasets with automated delineations (test sets of OASIS and MS dataset).

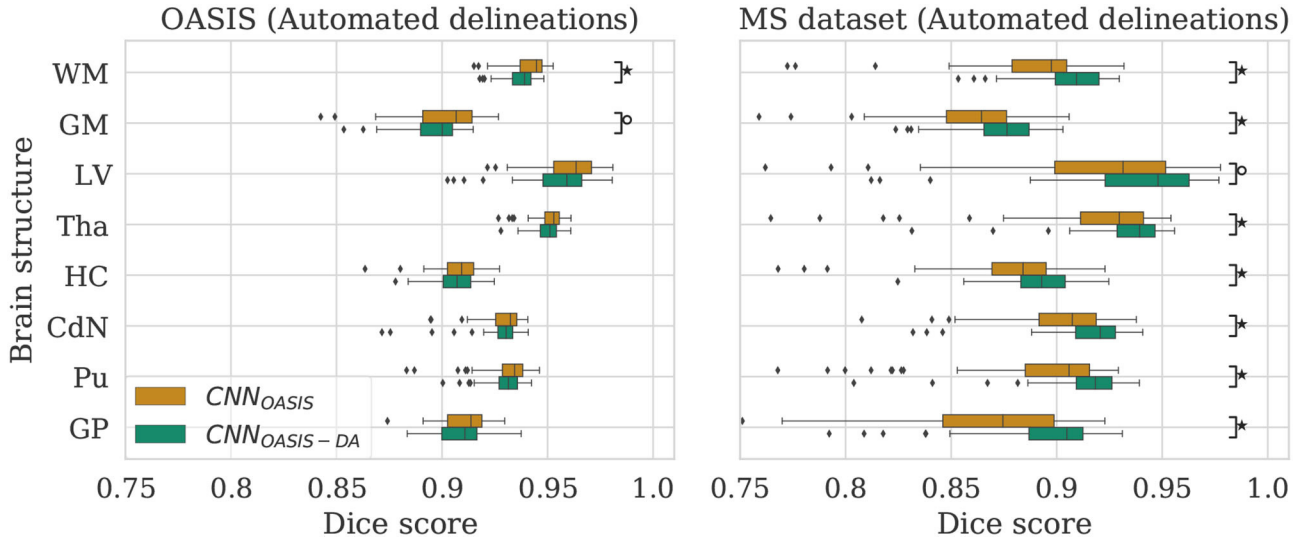


FIGURE 3 | Dice scores for the CNN_{OASIS} and $CNN_{OASIS-DA}$ models on the OASIS (left) and MS dataset (right) test sets. Marks indicate that there is a significant difference between the two models (Wilcoxon, $^{\circ}p < 0.05$, $^{\star}p < 5\times10^{-4}$).

TABLE 2 | Summary of the Dice score (DC) performance of models trained on the OASIS data (CNN_{OASIS} and $CNN_{OASIS-DA}$) and tested on the OASIS and MS dataset test sets.

	OASIS-test set		MS dataset-test set	
Tissues	**CNN_{OASIS}**	**$CNN_{OASIS-DA}$**	**CNN_{OASIS}**	**$CNN_{OASIS-DA}$**
WM	**0.945**	0.939	0.897	**0.909**
GM	**0.907**	0.900	0.864	**0.876**
LV	0.964	0.959	0.931	**0.948**
Tha	0.953	0.951	0.930	**0.939**
HC	0.909	0.907	0.884	**0.893**
CdN	0.932	0.930	0.907	**0.921**
Pu	0.934	0.931	0.906	**0.918**
GP	0.914	0.911	0.874	**0.905**
ALL	**0.932**	0.929	0.899	**0.914**

Highlighted results indicate that median values are larger (P50: Wilcoxon, $p < 0.05$).

This will allow us to determine: (i) if applying GMM-DA decreases the performance on data similar to the training set in comparison to the base model, and (ii) how the models perform in a multi-scanner setting. The results in terms of Dice scores are summarized in **Figure 3** and **Table 2**. The corresponding *Se* and *Pr* results can be found in the **Supplementary Table 1**.

4.1.1. OASIS

The models achieve high Dice scores and low variability. *Se* and *Pr* are very similar for CNN_{OASIS} and $CNN_{OASIS-DA}$ (min: $Se_{GM} = 0.87$, $Pr_{GM} = 0.87$; mean: $\overline{Se} = 0.94$, $\overline{Pr} = 0.94$). There is no statistical difference between the Dice score results (Wilcoxon: $p > 0.05$, Levene: $p > 0.05$), except for WM and GM, where CNN_{OASIS} tends to perform better (Wilcoxon, $p < 0.05$). Although statistically different, the difference is marginal, especially when considering the lower limits of the distributions, as can be appreciated on the left hand side panel of **Figure 3**.

4.1.2. MS Dataset

$CNN_{OASIS-DA}$ outperforms CNN_{OASIS} for all structures (Wilcoxon: $p \ll 0.05$). *Se* values are also lower in the CNN_{OASIS} model (min: $Se_{GP} = 0.81$, mean: $\overline{Se} = 0.88$), while *Pr* values are overall comparable between the two models, with local differences for specific tissues (refer to **Supplementary Table 1** for details). Additionally, we can observe in the right hand side panel of **Figure 3** that the variability and incidence of outliers is reduced for $CNN_{OASIS-DA}$. All these observations imply that the addition of GMM-DA greatly improves the performance of

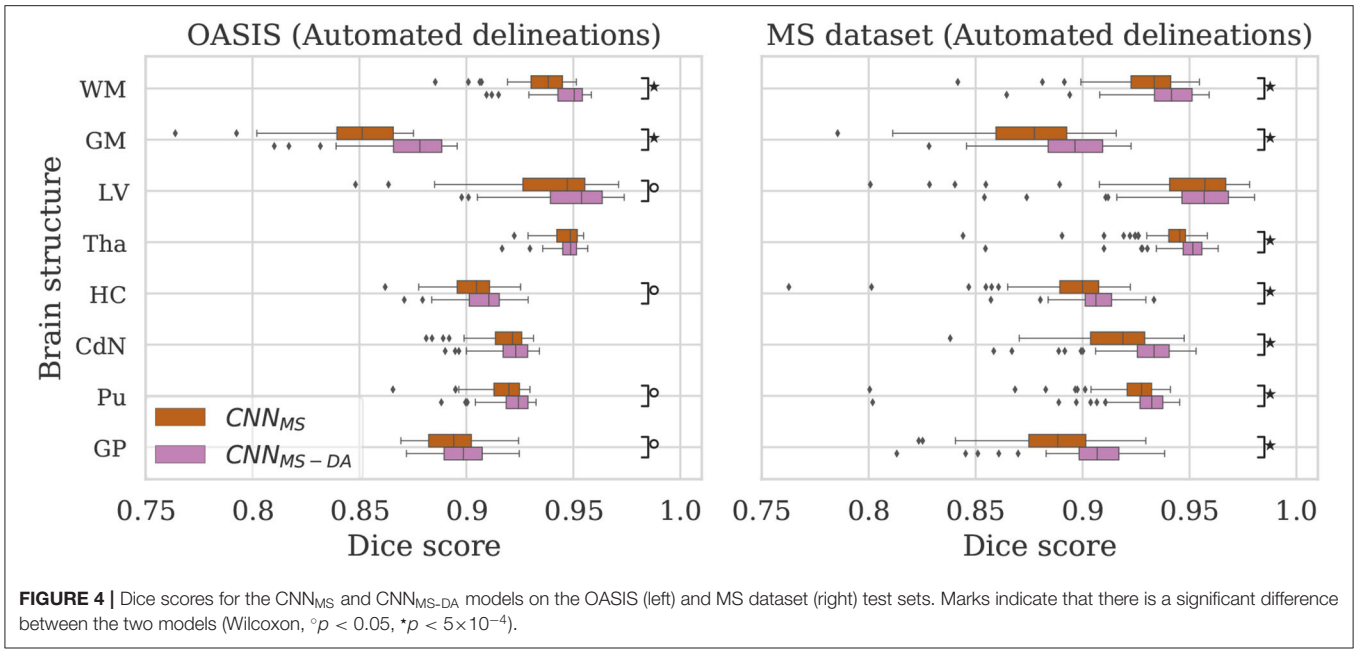


FIGURE 4 | Dice scores for the CNN_{MS} and $CNN_{MS\text{-}DA}$ models on the OASIS (left) and MS dataset (right) test sets. Marks indicate that there is a significant difference between the two models (Wilcoxon, $^{\circ}p < 0.05$, $^{\star}p < 5{\times}10^{-4}$).

TABLE 3 | Summary of the Dice score (DC) performance of models trained on the MS dataset (CNN_{MS} and $CNN_{MS\text{-}DA}$) and tested on the OASIS and MS dataset test sets.

	OASIS-test set		MS dataset-test set	
Tissues	**CNN_{MS}**	**$CNN_{MS\text{-}DA}$**	**CNN_{MS}**	**$CNN_{MS\text{-}DA}$**
WM	0.938	**0.950**	0.934	**0.942**
GM	0.851	**0.878**	0.877	**0.896**
LV	0.947	**0.954**	0.957	0.957
Tha	0.949	0.949	0.945	**0.952**
HC	0.905	**0.910**	0.900	**0.906**
CdN	0.922	0.923	0.919	**0.933**
Pu	0.920	**0.924**	0.928	**0.932**
GP	0.894	**0.899**	0.888	**0.907**
ALL	0.916	**0.923**	0.919	**0.928**

Highlighted results indicate that median values are larger (P50: Wilcoxon, $p < 0.05$).

the model to new data containing unseen scanner types from different centers.

4.2. GMM Augmentation of a Heterogeneous Dataset

Now that we have established that the addition of GMM-DA is beneficial for the generalization of a model trained on a homogeneous dataset to multi-scanner settings, we evaluate the performance of a model trained on the MS dataset, which is very heterogeneous. We additionally investigate the effect of adding GMM-DA when training on a dataset with these characteristics. The CNN_{MS} and $CNN_{MS\text{-}DA}$ models are evaluated in the same way as the above, and results are summarized in **Figure 4** and **Table 3**. The corresponding *Se* and *Pr* results can be found in the **Supplementary Table 2**.

4.2.1. OASIS

The MS dataset does not contain images with the same characteristics as OASIS. This explains a drop in performance in terms of DC for CNN_{MS} on the OASIS test set. From **Table 3** we can observe that after the addition of GMM-DA the performance increases: $CNN_{MS\text{-}DA}$ performs better for all the structures, with the exception of Tha and CdN, where there is no statistical difference in terms of performance (see the left hand side panel of **Figure 4**).

4.2.2. MS Dataset

As mentioned in section 3, the MS test set contains scanner types which were not present in the training set. $CNN_{MS\text{-}DA}$ outperforms CNN_{MS} for all structures (Wilcoxon: $p \ll 0.05$) except LV (Wilcoxon: $p > 0.05$) in terms of DC (see **Table 3** and the right hand side panel of **Figure 4**). *Se* and *Pr* values are

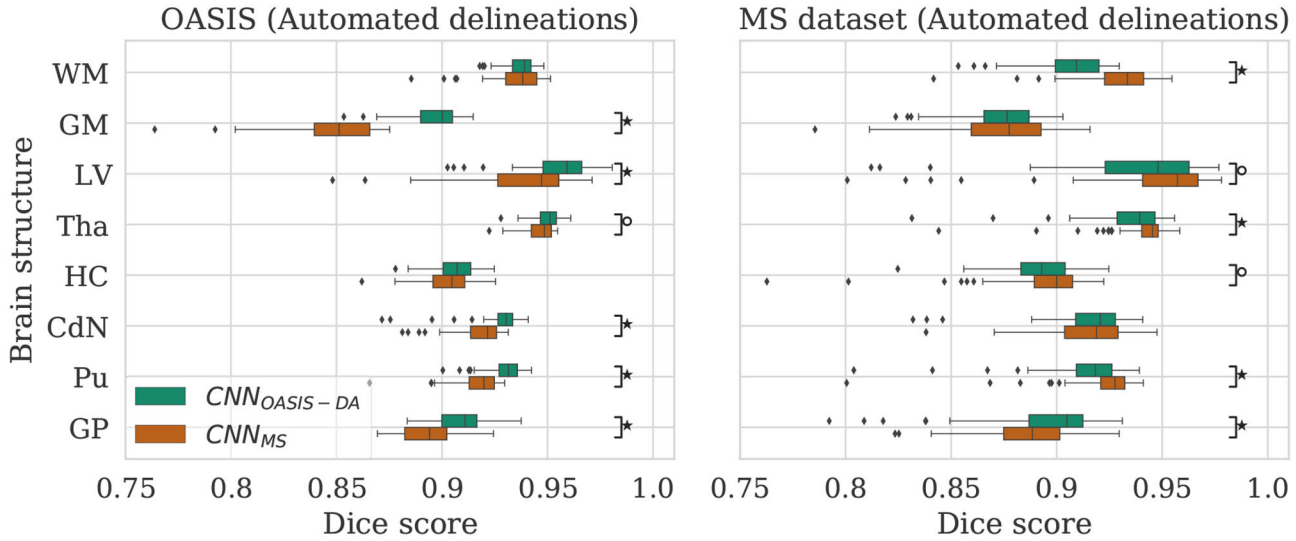


FIGURE 5 | Dice scores for the $CNN_{OASIS-DA}$ and CNN_{MS} models on the OASIS (left) and MS dataset (right) test sets. Marks indicate that there is a significant difference between the two models (Wilcoxon, $^{\circ}p < 0.05$, $^{\star}p < 5\times10^{-4}$).

also generally lower in the CNN_{MS} model, with local differences for specific tissues (see **Supplementary Table 2** for details). This indicates that adding GMM-DA to an already heterogeneous dataset can further increase the generalizability of the network.

4.3. Comparison Between the Different Models

Given that large multi-scanner and multi-center datasets are not commonly available to researchers, we are particularly interested in the comparison between the model trained on OASIS with augmentation ($CNN_{OASIS-DA}$) against the model trained on the MS dataset without augmentation (CNN_{MS}). To facilitate the comparison, the performance of both models is displayed in **Figure 5**.

4.3.1. OASIS

For most of the evaluated structures, CNN_{MS} shows a significant decrease in performance in comparison to $CNN_{OASIS-DA}$. It is expected that the models trained on the MS dataset have generally lower performance than the models trained on OASIS, since the images in the MS dataset training set do not share the same characteristics as the ones in OASIS (as previously illustrated in **Figure 2**). The addition of GMM-DA to CNN_{MS} can help reduce this performance gap, as seen in the previous section.

4.3.2. MS Dataset

Analyzing the right hand side panel of **Figure 5**, it is interesting to verify that $CNN_{OASIS-DA}$ approximates the variability of the CNN_{MS} for all the structures. In terms of median DC values it sometimes equals or even surpasses its performance (GM, GP and CdN). It is important to keep in mind that the MS dataset contains pathological images which are not present in OASIS. CNN_{MS} has been exposed to many more types of images, with some patients possibly presenting a small number of lesions. However, the contrary is not true, given that OASIS only contains images from healthy subjects. At best, the networks trained on this data were exposed to a few lesions present in the older subjects' scans. It is thus not possible to guarantee that the differences in performance between CNN_{MS} and $CNN_{OASIS-DA}$ on a pathological dataset are caused only by scanner or acquisition variability. Nevertheless, these results show that with a simple data augmentation strategy it is possible to achieve competitive results on unseen data from various scanners and centers.

In order to visualize the different results, **Figure 6** illustrates the results obtained on three different images from the MS dataset using the four different models described so far. For simplicity, WM and GM are not shown. Looking at this figure it is very clear that when the image contrast is not good, the CNN_{OASIS} model can produce segmentation results which infiltrate WM and CGM regions in unexpected ways. The addition of GMM-DA brings the results much closer to the ground truth results.

4.4. Evaluation on Manual Labels

To validate the performance of the models on manual segmentations we evaluate them on the MICCAI 2012 dataset. It is interesting to compare their performance against the performance of the method used to get the automated labels the models were trained on (ico**brain**). The results are summarized in **Figure 7**, where results which are statistically different to ico**brain** are indicated (Wilcoxon: $p < 0.05$).

For most structures the models reach comparable performance. CNN_{MS-DA} is the model with overall best performance, but still does not surpass ico**brain**. For GM, CNN_{MS} and CNN_{MS-DA} achieve much lower performance than the other models. This is in line with the results observed for the OASIS dataset. Recalling that this dataset is derived from a

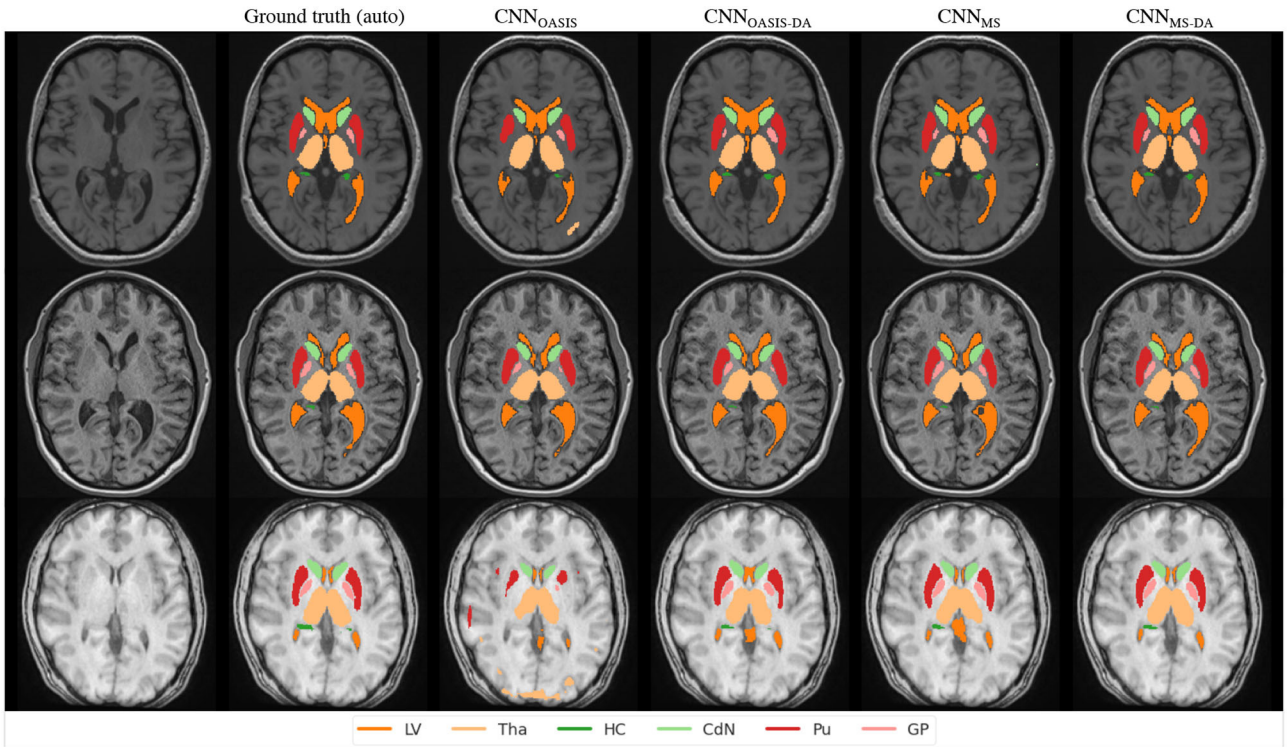


FIGURE 6 | Segmentation results for the different models on three test examples from the MS dataset. The automated ground truth is computed by ico**brain**. Results from CNN$_{OASIS}$ are very variable depending on image intensity. The addition of GMM-DA improves the segmentation prediction.

subset of OASIS, CNN$_{OASIS}$ and CNN$_{OASIS-DA}$ were exposed to images with these characteristics during training, while CNN$_{MS}$ and CNN$_{MS-DA}$ were not. Variances are not statistically different for any tissue type. *Se* and *Pr* values are also comparable for all models, with mean $\overline{Se} \approx 0.84, \overline{Pr} \approx 0.85$.

4.5. Consistency on *Test-Retest* Data

By evaluating the models on the test-retest dataset described in section 3 it is possible to evaluate how each model deals with differences in scanner type. As previously mentioned, the dataset contains two repetitions per scanner in two or three different scanners. We compute the difference in predicted volume for each of the evaluated structures between same scanner repetitions (*intra-scanner* differences) and between the repetitions in different scanners (*inter-scanner* differences). We consider all possible scanner combinations, which means that we end up with 26 intra-scanner and 88 inter-scanner repetitions. We compare the performance of our methods against ico**brain**. As already mentioned, this method is clinically available. However, when performing longitudinal evaluations, this method has a key limitation: the results are considered reliable only if the two images being analyzed were acquired in the same, or compatible, scanner. As such, we are interested in achieving better inter-scanner volume estimation differences, and we consider inter-scanner results to be consistent if the volume differences are in a comparable range to the intra-scanner differences obtained by ico**brain**.

For a simplified overview of the results, we plot the distribution of volume differences for all the considered brain structures in **Figure 8**. Additionally, in **Table 4** we showcase the results in terms of median and *P*90, which translates the variability in the distributions. We exclude the CNN$_{OASIS}$ model from the table, since it is clear from **Figure 8** and **Table 2** that the performance of this method is low for multi-scanner datasets.

Globally we observe that intra-scanner differences are much lower than inter-scanner differences for all the models. In the intra-scanner case, CNN$_{OASIS}$ produces a higher error than the other models for all structures. Interestingly, CNN$_{OASIS-DA}$ produces very stable results, comparable to or even better than ico**brain** for several structure types (Tha, HC, CdN). CNN$_{MS-DA}$ produces the most consistent results for most of the structures, especially when considering *P*90.

Regarding inter-scanner differences, we observe that the CNN$_{OASIS}$ model produces extremely large variability. The other models either compare to ico**brain** or produce more consistent results. The exception is WM and GM, where ico**brain** still outperforms the other methods in terms of consistency. This is in line with the previous observations that performance (in terms of Dice) was lower in these two tissues. The most important observation is that CNN$_{MS-DA}$ produces the most consistent results for all the substructures. The results for this model are

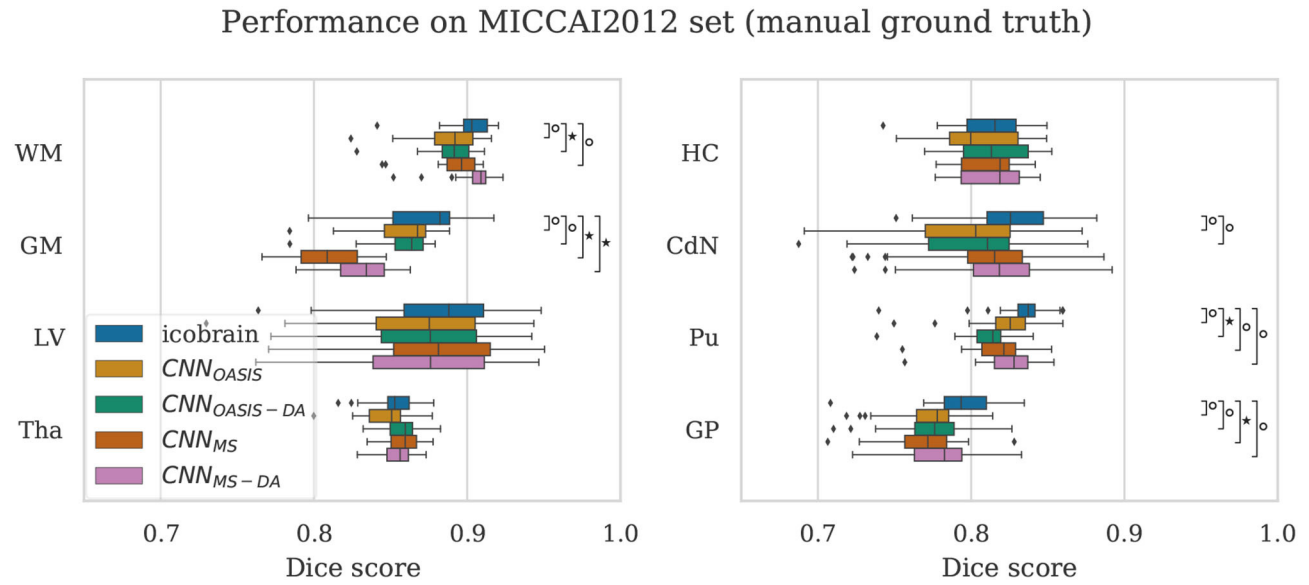


FIGURE 7 | Dice scores on the MICCAI2012 test set for ico**brain**, CNN_{OASIS}, $CNN_{OASIS-DA}$, CNN_{MS} and CNN_{MS-DA} models. Asterisks indicate that there is a significant difference between the result and ico**brain** (Wilcoxon, $^{\circ}p < 0.05$, $^{\star}p < 5\times10^{-4}$).

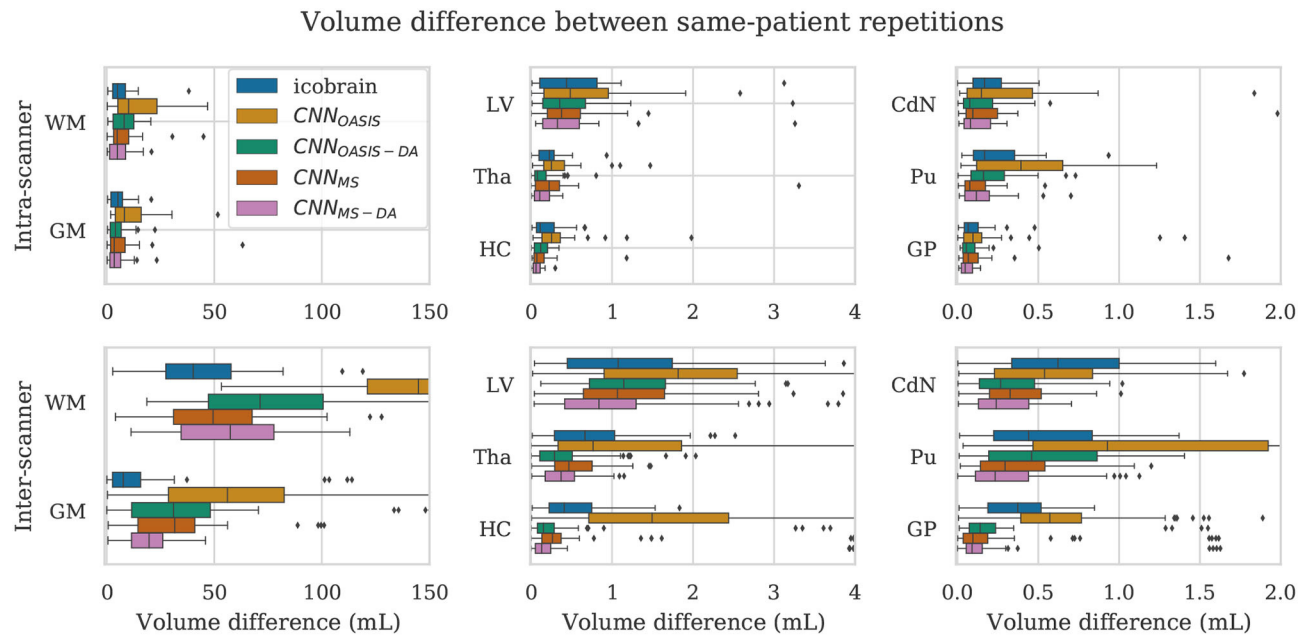


FIGURE 8 | Volume difference between intra-scanner and inter-scanner repetitions from the same patient. To be able to keep the same scale for intra-scanner and inter-scanner the CNN_{OASIS} results are sometimes not fully shown.

sometimes comparable to the values obtained by ico**brain** in the intra-scanner case (noticeably for HC and GP). Overall, the addition of GMM-DA results in a very significant improvement, both in comparison to ico**brain** and to the CNN_{MS} method. Additionally, a very interesting observation is that $CNN_{OASIS-DA}$ achieves a performance which is comparable to that of CNN_{MS}, sometimes even surpassing it (Tha, HC, CdN).

4.6. Influence of Bias Field

A bias field is an undesirable spatially smoothly varying low frequency signal that often corrupts MRI images (Juntu et al., 2005). A number of methods have been proposed to remove this signal from the images, and bias-field correction is often used as a pre-processing step. Given that this is a slow procedure which can sometimes produce underlying errors, it has become

TABLE 4 | Volume differences (mL) between intra- and inter-scanner repetitions from the same patient.

		Intra-scanner differences							
		WM	**GM**	**LV**	**Tha**	**HC**	**CdN**	**Pu**	**GP**
ico**brain**	P50	5.00	5.22	0.44	0.23	0.11	0.17	0.17	0.07
	P90	**12.94**	12.38	1.04	0.47	0.56	0.43	0.48	0.19
$CNN_{OASIS-DA}$	P50	8.08	4.06	0.35	**0.08**	0.12	**0.08**	0.16	0.06
	P90	19.72	13.54	1.17	0.43	0.31	0.34	0.44	0.16
CNN_{MS}	P50	**4.86**	**3.48**	0.38	0.22	0.07	0.09	**0.08**	0.07
	P90	15.82	14.73	1.02	0.46	0.28	0.36	**0.28**	0.19
CNN_{MS-DA}	P50	4.94	3.51	**0.33**	0.11	**0.06**	**0.08**	0.12	**0.05**
	P90	14.11	**12.12**	**0.82**	**0.33**	**0.16**	**0.26**	0.34	**0.13**
		Inter-scanner differences							
		WM	**GM**	**LV**	**Tha**	**HC**	**CdN**	**Pu**	**GP**
ico**brain**	P50	**40.26**	**7.68**	1.08	0.67	0.41	0.62	0.44	0.37
	P90	**76.52**	**23.77**	2.73	1.38	1.15	1.16	1.04	0.67
$CNN_{OASIS-DA}$	P50	71.28	31.07	1.15	**0.29**	0.15	0.27	0.46	0.14
	P90	117.65	65.34	3.15	1.21	0.56	0.82	1.11	0.30
CNN_{MS}	P50	49.43	31.58	1.07	0.47	0.26	0.33	0.29	0.10
	P90	94.36	55.98	**2.56**	1.46	0.65	0.91	1.13	0.73
CNN_{MS-DA}	P50	57.45	19.71	**0.84**	0.37	**0.13**	**0.24**	**0.23**	**0.09**
	P90	93.47	35.75	2.85	**0.88**	**0.29**	**0.57**	**0.87**	**0.27**

Best results are highlighted.

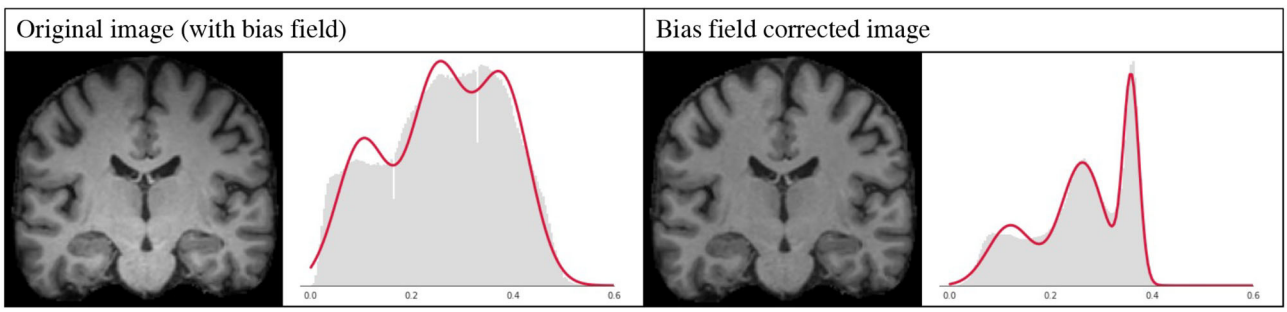


FIGURE 9 | Influence of bias field on the GMM. Images and corresponding histograms with GMM prediction overlaid in red. Left: original image with bias field; right: bias field corrected image.

popular to skip bias field correction when using deep learning approaches, and instead allow the networks to learn the bias-field mechanisms, with good results (Kamnitsas et al., 2017).

However, bias field correction is extremely important for GMM-based methods, since it changes the intensity profiles of the different tissues. This effect is illustrated in **Figure 9**, where the histogram of a bias field corrected image is compared to that of an image with bias field. It is very likely that when applying GMM-DA some of the voxels corresponding to WM will be treated as GM, or vice-versa. This implies that the structural information can be lost, which will very likely result in drop in the performance of a model trained on images with bias field and the addition of GMM-DA.

To test our hypothesis we trained two models, following the same scheme as in the previous experiments, on images with bias field. More specifically, we repeated the experiment from the previous section training on the MS dataset without the bias field correction at pre-processing. The results of this experiment are detailed in **Table 5**. As expected, applying GMM-DA on this type of data either decreases the performance of the method (WM, GM and LV), or has no effect on the segmentation performance. This is in line with our hypothesis

TABLE 5 | Summary of the performance of models trained and evaluated on data with bias field on MS dataset.

	MS dataset-test set (with bias field)			
	$CNN_{MS\text{-}BF}$		$CNN_{MS\text{-}BF\text{-}DA}$	
Tissue	**DC(P50)**	**DC(P10)**	**DC(P50)**	**DC(P10)**
WM	**0.950**	0.920	0.945	0.914
GM	**0.914**	0.877	0.903	0.859
LV	**0.941**	**0.854**	0.845	0.554
Tha	0.937	0.911	0.937	0.896
HC	0.865	0.842	0.858	0.827
CdN	0.915	0.880	0.910	0.865
Pu	0.903	0.885	0.911	0.891
GP	0.855	0.803	0.854	0.807
ALL	0.910	0.872	0.895	0.827

DC, Dice scores. Highlighted results indicate that median values are larger (P50: Wilcoxon, $p < 0.05$) or variances are lower (P10: Levene, $p < 0.05$).

and indicates that the GMM-DA should be applied on bias-field corrected images.

5. CONCLUSIONS AND FUTURE WORK

In this work we present a novel intensity-based data augmentation strategy. The main goal of this approach is to aid models trained on scanner- and center-homogeneous datasets generalizing to multi-scanner, multi-center data. The proposed method is fast, simple and can be added to any MRI training pipeline to generate images on-the-fly. We observed that applying the augmentation step while training on homogeneous data leads to a pronounced improvement in performance when the trained model is tested in multi-scanner data from difference centers. This is the case in terms of segmentation quality (as measured by Dice score), but also in the consistency of the produced prediction (as measured in terms of volume differences). When applied to the test-retest dataset there is a remarkable improvement, especially for repetitions in different scanners. The baseline model trained on homogeneous data produces extremely inconsistent results, while the same model with addition of GMM-DA compares to a model trained on multi-scanner, multi-center data. We additionally verify that applying GMM-DA when training a model on multi-center data results in an increase in performance, again both in terms of accuracy and consistency of the predictions. These observations are particularly interesting because large multi-scanner, multi-center datasets are not commonly available to researchers in the field. Nevertheless, even when such a dataset is available, it is possible to obtain even more generalization by adding a simple augmentation strategy.

It should be noted that the heterogeneous dataset contains several sources of variability, including acquisition sequence parameters. The resulting contrast variability is also addressed by the GMM-DA. Therefore, we can attribute the improvement in the generalization capabilities of the CNN not only to scanner, but also to generalization to unseen acquisition parameters, or other center-specific factors.

It is possible that combining this method with other DA procedures would result in an even more robust model. Nevertheless, we opted to restrict the augmentation procedures such that we could observe the added value of our method alone. Additionally, since the images were registered to MNI space adding geometric transformations such as rotations and flips is not necessary. Nonetheless, it is expected that the DA algorithm still works well if the images are in native space. Registration was performed as a way to simplify the learning of the network, since we were interested in comparing the effect of the augmentation step in a simplified setting.

There are a few limitations to the present work. Namely, the images need to be bias-field corrected as a pre-processing step to successfully apply the GMM-DA. We don't see this as a disadvantage, since GMM-DA is only needed at training time. We argue that it would be possible to add back the bias-field to the augmented image, which would allow the model to be effectively trained with bias field. This step would allow the final trained model to generalize to images with bias-field, thus eliminating the need for bias-field correction at inference time. Experimental validation of this claim remains out of the scope of the present work, given that it is related to improving the overall model performance, and is not connected to the effectiveness of the proposed approach.

Additionally, the presence of pathology in the MS dataset introduces an extra source of variability. In images with WM lesions, as is the case for MS, it is tempting to assume that a fourth component to the GMM would be a good way to capture the lesion class. However, lesions in T1w images overlap with the GM class in terms of intensity, for which reason it would be impossible to perfectly disentangle the two classes with the current framework. A more sophisticated approach would be necessary for this, likely at the cost of the possibility to generate images on-the-fly, unless lesion masks are available.

Finally, due to scarcity of manual delineations, the models were trained on automated segmentations. This is not ideal, because our model is likely to inherit any bias or known problems that might exist in the ground truth. However, given that we are

especially interested in the effect of the augmentation we can still make a fair comparison between the approaches.

Although we focused on the task of brain structure segmentation in T1w MRI images, we believe this simple method has the potential to be used for other tasks in medical imaging that make use of MR images. As long as there are discernible, anatomically-related peaks in the intensity histograms, the method is transferable to other MR protocols and sequences. It is an open question whether the method is helpful for different tasks without further adaptations. For tasks such as lesion segmentation we hypothesize that if lesion masks are available it would be simple to adapt the method such that contrasts and intensities are locally modified within the abnormal area. We further see potential in this method to be adapted such that it offers a fast way to replace missing modalities in tasks requiring two or more MRI modalities (e.g., as often performed for brain tumor segmentation). This would expectably come at the expense of some performance power, but could allow existing pipelines to be used on incomplete data.

Given these considerations, an immediate next step would be to apply the current method to different applications (e.g., brain age or disability scores prediction from MR images) and verify our claim. A second step would be to extend the method to different types of brain lesions when such masks are available, to model the intensity of the tissues of interest individually, and test the added value of the extended method to applications such as detection, classification and segmentation of MS lesions, stroke, or brain tumors. Additional future directions include extending the augmentation method by introducing changes to the different components of the mixture such that they are not necessarily represented by Gaussian distributions. Moreover, it would be interesting to investigate how the addition of (preferably Rician) noise to the images would impact performance on unseen scanner types. Typical geometric distortions and bias fields can also be modeled and included in a more complex data augmentation scheme.

DATA AVAILABILITY STATEMENT

This research study was conducted using human subject data partly made available in open access by OASIS (https://www.oasis-brains.org/) (Marcus et al., 2007) and manual labelings by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) under academic subscription (Landman and Warfield, 2012). The data is released under the Creative Commons Attribution NonCommercial license (CC BY-NC) with no end date. The MS dataset is a subset of data processed with icobrain ms in clinical practice, for which subjects had agreed to allow icometrix to use an anonymized version of the already analysed MR images for research purposes.

AUTHOR CONTRIBUTIONS

MM: conceptualization, investigation, methodology, software, and writing—original draft preparation. ER and RP: methodology and writing—reviewing and editing. NP: conceptualization and writing—reviewing and editing. KV: supervision, writing—reviewing and editing, and funding acquisition. DS: supervision, conceptualization, writing—reviewing and editing, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreements nos 765148 and 764513, by the NIH NINDS grant no R01NS112161 and by the Penta project 19021 (Vivaldy).

ACKNOWLEDGMENTS

This work builds on the preliminary research paper *An augmentation strategy to mimic multi-scanner variability in MRI*, accepted for presentation at the International Symposium on Biomedical Imaging (ISBI) 2021 and to be published in conference proceedings (Meyer et al., 2021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.708196/full#supplementary-material

REFERENCES

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Bagc,i, U., Udupa, J. K., and Bai, L. (2010). The role of intensity standardization in medical image registration. *Pattern Recognit. Lett*. 31, 315–323. doi: 10.1016/j.patrec.2009.09.010

Billot, B., Greve, D. N., Van Leemput, K., Fischl, B., Iglesias, J. E., and Dalca, A. (2020). "A learning strategy for contrast-agnostic mri segmentation," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning, Vol. 121 of Proceedings of Machine Learning Research*, (Montreal, QC: PMLR), 75–93.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016* (Cham: Springer International Publishing), 424–432.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., et al. (2019). DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* 64m 160–170. doi: 10.1016/j.mri.2019.05.041

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167, 104–120. doi: 10.1016/j.neuroimage.2017.11.024

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., and Shinohara, R. T. (2016). Removing inter-subject technical variability

in magnetic resonance imaging studies. *Neuroimage* 132, 198–212. doi: 10.1016/j.neuroimage.2016.02.036

Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W. H., Corvin, A., et al. (2020). Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *Neuroimage* 220, 117127. doi: 10.1016/j.neuroimage.2020.117127

Giorgio, A., and De Stefano, N. (2013). Clinical use of brain volumetry. *J. Magn. Reson. Imaging* 37, 1–14. doi: 10.1002/jmri.23671

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Jain, S., Sima, D. M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., et al. (2015). Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage Clin.* 8:367–375. doi: 10.1016/j.nicl.2015.05.003

Jog, A., Hoopes, A., Greve, D. N., Van Leemput, K., and Fischl, B. (2019). PSACNN: Pulse sequence adaptive fast whole brain segmentation. *Neuroimage* 199, 553–569. doi: 10.1016/j.neuroimage.2019.05.033

Juntu, J., Sijbers, J., Van Dyck, D., and Gielen, J. (2005). "Bias field correction for mri images," in *Computer Recognition Systems* (Berlin; Heidelberg: Springer Berlin Heidelberg), 543–551.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36:61–78. doi: 10.1016/j.media.2016.10.004

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems, Vol. 25*. (Lake Tahoe, NV: Curran Associates, Inc).

Landman, B., and Warfield, S. (2012). "Miccai 2012 workshop on multi-atlas labeling," in *MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling* (Nice: CreateSpace Independent Publishing Platform).

Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., et al. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* 66:101714. doi: 10.1016/j.media.2020.101714

Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). *Rectifier Nonlinearities Improve Neural Network Acoustic Models*. Technical report.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* 19, 1498–1507. doi: 10.1162/jocn.2007.19.9.1498

Meyer, M. I., de la Rosa, E., Barros, N., Paolella, R., Van Leemput, K., and Sima, D. M. (2021). "An augmentation strategy to mimic multi-scanner variability in MRI," in *2021 IEEE 18th IEEE International Symposium* on *Biomedical Imaging* (Nice). 1196–1200.

Meyer, M. I., de la Rosa, E., Van Leemput, K., and Sima, D. M. (2019). "Relevance vector machines for harmonization of MRI brain volumes using image descriptors," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes Bioinformatics)*, *Vol. 11796 LNCS* (Shenzhen), 77–85.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford University, CA), 565–571.

Moyer, D., and Golland, P. (2021). Harmonization and the worst scanner syndrome. *arXiv, cs.LG/2101.06255*.

Nyúl, L. G., and Udupa, J. K. (1999). On standardizing the MR image intensity scale. *Magn. Reson. Med.* 42, 1072–1081.

Ourselin, S., Roche, A., Subsol, G., Pennec, X., and Ayache, N. (2001). Reconstructing a 3d structure from serial histological sections. *Image Vis. Comput.* 19, 25–31. doi: 10.1016/S0262-8856(00)00052-4

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015* (Cham: Springer International Publishing), 234–241.

Salimans, T., and Kingma, D. P. (2016). "Weight normalization: a simple reparameterization to accelerate training of deep neural networks," in *Conference on Neural Information Processing Systems* (Barcelona), 901–909.

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., et al. (2018). "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging* (Cham: Springer International Publishing), 1–11.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin.* 6, 9–19. doi: 10.1016/j.nicl.2014.08.008

Sietsma, J., and Dow, R. J. (1991). Creating artificial neural networks that generalize. *Neural Networks* 4, 67–79. doi: 10.1016/0893-6080(91)90033-2

Simard, P., Steinkraus, D., and Platt, J. (2003). "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings* (Edinburgh), 958–963.

Struyfs, H., Sima, D. M., Wittens, M., Ribbens, A., Pedrosa de Barros, N., Phan, T. V., et al. (2020). Automated MRI volumetry as a diagnostic tool for Alzheimer's disease: validation of icobrain dm. *Neuroimage Clin.* 26:102243. doi: 10.1016/j.nicl.2020.102243

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Cham: Springer International Publishing), 240–248.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4itk: Improved n3 bias correction. *IEEE Trans Med Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated model-based tissue classification of MR images of the brain. *IEEE Trans. Med. Imaging* 18, 897–908. doi: 10.1109/42.811270

Wang, L., Lai, H. M., Barker, G. J., Miller, D. H., and Tofts, P. S. (1998). Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. *Magn. Reson. Med.* 39, 322–327. doi: 10.1002/mrm.1910390222

Wrobel, J., Martin, M. L., Bakshi, R., Calabresi, P. A., Elliot, M., Roalf, D., et al. (2020). Intensity warping for multisite MRI harmonization. *Neuroimage* 223:117242. doi: 10.1016/j.neuroimage.2020.117242

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., and Dalca, A. V. (2019a). "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA).

Zhao, F., Wu, Z., Wang, L., Lin, W., Xia, S., Shen, D., et al. (2019b). "Harmonization of infant cortical thickness using surface-to-surface cycle-consistent adversarial networks," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes Bioinformatics)*, *Vol. 11767 LNCS*, 475–483.

Zhuge, Y., and Udupa, J. K. (2009). Intensity standardization simplifies brain MR image segmentation. *Comput. Vis. Image Understand.* 113, 1095–1103. doi: 10.1016/j.cviu.2009.06.003

Conflict of Interest: MM, ER, NP, RP, and DS are employed by ico**metrix**.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

HYPOTHESIS AND THEORY
published: 15 November 2021
doi: 10.3389/fnins.2021.595412

Altered Spontaneous Brain Activity Related to Neurologic and Sleep Dysfunction in Children With Obstructive Sleep Apnea Syndrome

Jie Bai[1†], Hongwei Wen[2,3†], Jun Tai[4], Yun Peng[1], Hongbin Li[5], Lin Mei[5], Tingting Ji[5], Xiaodan Li[5], Yanhua Li[1], Xin Ni[5] and Yue Liu[1]**

[1] Department of Radiology, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China, [2] Key Laboratory of Cognition and Personality, Ministry of Education, Chongqing, China, [3] School of Psychology, Southwest University, Chongqing, China, [4] Department of Otolaryngology, Head and Neck Surgery, Children's Hospital Capital Institute of Pediatrics, Beijing, China, [5] Department of Otolaryngology, Head and Neck Surgery, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China

OPEN ACCESS

Edited by:
Tim B. Dyrby,
Technical University of Denmark, Denmark

Reviewed by:
Xize Jia,
Hangzhou Normal University, China
Zhiliang Long,
Southwest University, China

***Correspondence:**
Yue Liu
liuyue20136@163.com
Xin Ni
nixin@bch.com.cn

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 16 August 2020
Accepted: 27 September 2021
Published: 15 November 2021

Citation:
Bai J, Wen H, Tai J, Peng Y, Li H, Mei L, Ji T, Li X, Li Y, Ni X and Liu Y (2021) Altered Spontaneous Brain Activity Related to Neurologic and Sleep Dysfunction in Children With Obstructive Sleep Apnea Syndrome. Front. Neurosci. 15:595412. doi: 10.3389/fnins.2021.595412

Childhood obstructive sleep apnea (OSA) is a common chronic sleep-related breathing disorder in children, which leads to growth retardation, neurocognitive impairments, and serious complications. Considering the previous studies about brain structural abnormalities in OSA, in the present study, we aimed to explore the altered spontaneous brain activity among OSA patients, using amplitude of low-frequency fluctuation (ALFF), fractional ALFF (fALFF), and regional homogeneity (ReHo) methods based on resting-state functional magnetic resonance imaging (MRI). Thirty-one untreated OSA children and 33 age-and gender-matched healthy children (HC) were included in this study. Compared with controls, the OSA group showed significant lower ALFF in the right lingual gyrus, decreased fALFF in the left middle frontal gyrus (MFG), but increased fALFF in the left precuneus. Decreased ReHo was found in the left inferior frontal gyrus (orbital part) and left middle frontal gyrus. Notably, the mean fALFF value of left MFG was not only significantly related to multiple sleep parameters but also demonstrated the best performance in ROC curve analysis. These findings revealed OSA children were associated with dysfunctions in the default mode network, the frontal lobe, and the lingual gyrus, which may implicate the underlying neurophysiological mechanisms of intrinsic brain activity. The correlation between the altered spontaneous neuronal activity and the clinical index provides early useful diagnostic biomarkers for OSA children as well.

Keywords: obstructive sleep apnea, children, amplitude of low-frequency fluctuation fractional, regional homogeneity, functional magnetic resonance imaging, resting state, spontaneous activity, cognitive impairment

INTRODUCTION

Obstructive sleep apnea (OSA) is a common chronic sleep-related breathing disorder characterized by partial or complete closure of the upper airway during sleep, which results in recurrent intermittent hypoxia, carbon dioxide retention, and frequent awakening (Chen et al., 2016; Stevens et al., 2020). Nowadays, prevalence of OSA increased strikingly with the increasing population age

and obesity. The total prevalence of OSA among children ranges from 1 to 4% (Kirk et al., 2017). Clinically, pediatric OSA always presents with loud and irregular snoring and sleep disorders, in contrast to adults, with a range of different symptoms including growth retardation, enuresis, and behavioral problems, such as attention deficit/hyperactivity (ADHD) disorders (Krysta et al., 2017). Besides, noteworthy neurocognitive impairments also occur in pediatric OSA, involving learning, memory, executive function, visuospatial function, and psychomotor development (Copes and Rosentswieg, 1972; Hamasaki et al., 2007; Krysta et al., 2017), which severely diminished academic performance, social adaptation, and even the quality of life in children. Nevertheless, compared with adults, the scholars have paid little attention to brain abnormalities of pediatric OSA before.

Several previous studies exhibited that cortical thinning or gray matter volume reductions occurred in the superior frontal, ventral medial prefrontal, the superior and lateral parietal cortices, the superior temporal lobe, and the brainstem in OSA children, which arose from delayed neuronal development, damage, or atrophy (Philby et al., 2017; Macey et al., 2018). However, unlike in adults, pediatric OSA observed cortical thickening, including the precentral gyrus, the mid-to-posterior insular cortices, and the right anterior insula cortex in OSA children, which might involve in hypoxia-induced inflammatory changes (Macey et al., 2018). Furthermore, entropy measures based on high-resolution T1-weighted imaging identified early damages of brain tissue integrity in pediatric OSA. The affected brain sites included the prefrontal cortex; corpus callosum; and insular, frontal, temporal, hippocampus, and cerebellar areas, which localized within regulated autonomic, respiratory, and cognitive functions (Kheirandish-Gozal et al., 2018). These findings indicated that pediatric OSA presented extensive injury to the brain structures, which might account for underlying executive and cognitive deficits in children.

Currently, resting-state functional magnetic resonance imaging (rs-fMRI) has been found to be a useful and non-invasive technique for detecting spontaneous neural activity. Using rs-fMRI, several studies have demonstrated that OSA patients have shown noteworthy resting-state functional connectivity (rsFC) deficits, especially in the default mode network (DMN), which plays a role in sustaining brain function in the resting-state (Prilipko et al., 2011; Chen et al., 2018). Zhang et al. (2013) reported that OSA patients showed significantly reduced rsFC within the anterior DMN and bilateral fronto-parietal network but increased rsFC between the posterior cingulate cortex and precuneus within the DMN. Similarly, decreased rsFC was observed in DMN subregions, including the medial prefrontal cortex, anterior cingulate, and posterior cingulate in patients with OSA (Chen et al., 2016). Li et al. (2016) found that patients with OSA displayed a dysfunction of rsFC between the right hippocampus formation and posterior cingulate cortex within the DMN and significant negative correlation with delayed memory. In addition, the frontal lobe as a crucial brain area involved in multiple cognitive functions has been increasingly reported. Yu H. et al. (2019) showed that OSA patients showed significantly increased rsFC between the left dorsal amygdala, the right ventrolateral amygdala, and the left inferior frontal gyrus. Regarding these findings, OSA may be related to the abnormal rsFC between distinct brain areas in DMN, fronto-parietal, and limbic system, whereas which area is more responsible for the observed abnormal connectivity was still unclear. It is meaningful to directly locate the abnormal regional spontaneous neural activity in OSA patients during resting state and their relationships with behavioral performances.

Amplitude of low-frequency fluctuation (ALFF), fractional ALFF (fALFF), and regional homogeneity (ReHo) are three major data-driven measures for quantification of spontaneous neural activity based on BOLD signals. ALFF detects the total power within the range between 0.01 and 0.10 Hz and positively correlates to the alterations of spontaneous neural activity (Zang et al., 2007; Zou et al., 2008). Moreover, fALFF measures the ratio of the specific power spectrum of low frequency to that of the total power in the entire frequency range (Qiu et al., 2019). Both ALFF and fALFF have been proven to exhibit greater test–retest reliability, especially in gray matter (Zuo et al., 2010). Notably, fALFF produces better effects in reducing the physiological noise than ALFF, and it can effectively suppress artifacts in non-specific brain regions, such as the ventricles and the vicinity of blood vessels (Zou et al., 2008). ReHo is a data-driven measure for the local measurement of spontaneous neural activity (Xia et al., 2018), and it can effectively evaluate resting-state brain activity based on the hypothesis that brain activity is more likely to occur in clusters rather than in a single voxel (Zhang et al., 2012). Recently, these methods have wide access to explore brain diseases with potential functional alterations, such as depression (Yu Y. et al., 2019), Alzheimer's disease (Cheng et al., 2019), Tourette syndrome (Liu et al., 2017), and so on. Accordingly, the combination of the three may provide more detailed information about the intrinsic activity than each method alone.

In the present study, we not only investigated abnormal intensity of neural activity via ALFF/fALFF analysis but also investigated abnormal neural synchronization via ReHo analysis in OSA children. Based on previous studies, we hypothesized that (1) OSA children would show altered ALFF/fALFF and ReHo values in the DMN, frontal lobe, and lingual gyrus; (2) the alterations of the spontaneous brain activity would be related to sleep-related respiratory parameters in OSA children; and (3) abnormal spontaneous activity pattern might be utilized as diagnostic neuroimaging biomarkers to discriminate OSA from controls. We aim to take a crucial step towards identifying spontaneous brain activity abnormalities in OSA children and providing potential targets for better understanding and treatment of this neurologic and sleep dysfunction disorder.

MATERIALS AND METHODS

Subjects

Thirty-one OSA children (age: 5.65 ± 2.82 years, range: 3–10 years; 12 female) were recruited in Beijing Children's Hospital from April 2016 to October 2019. We also included 33 age- (p = 0.585, two-sample t-test) and gender-matched (p = 0.431, chi-square test) healthy children (age: 6.01 ± 2.43 years; range: 2–11 years; 16 female) in our study (**Table 1**). All the participants

TABLE 1 | Demographic and clinical characteristics of OSA patients and healthy controls.

Characteristics		OSA (n = 31)	HC (n = 33)	p-value
Gender		19M/12F	17M/16F	$0.431^{\chi 2}$
Age (y)		5.65 ± 2.82	6.01 ± 2.43	0.585^{t}
Weight (kg)		29.14 ± 18.49	30.05 ± 13.59	0.842^{t}
Duration of disease (y)		1.79 ± 1.07	–	–
BMI (kg/mm^2)		18.39 ± 5.05	18.40 ± 3.33	0.996^{t}
AHI (per hour)		12.90 ± 13.89	1.34 ± 1.30	$<0.001^{t}$
OAI (per hour)		2.30 ± 3.47	0.07 ± 0.15	0.002^{t}
HI (per hour)		8.31 ± 11.05	0.69 ± 0.77	0.001^{t}
$LSaO_2$ (%)		87.82 ± 6.76	93.49 ± 2.84	$<0.001^{t}$
SaO_2 < 90% (%)		1.31 ± 2.45	0.00 ± 0.00	0.002^{t}
Sleep efficiency (%)		83.02 ± 10.28	89.03 ± 8.91	0.016^{t}
AI (per hour)		4.89 ± 6.26	0.74 ± 0.83	0.002^{t}
WISC-V	IQ	97.08 ± 7.69	103.27 ± 15.23	0.229^{t}
	VIQ	96.00 ± 6.90	106.09 ± 14.75	0.052^{t}
	PIQ	98.38 ± 10.25	99.18 ± 16.42	0.869^{t}
SAFE		9.80 ± 0.41	9.73 ± 0.47	0.656^{t}
Attention		17.08 ± 6.50	13.54 ± 9.24	0.233^{t}
Gesell	Adaptability DQ	80.86 ± 6.15	–	–
Developmental	GMQ	88.14 ± 6.34	–	–
Scale	FMQ	95.43 ± 8.98	–	–
	Language DQ	85.29 ± 11.32	–	–
	Personal-social DQ	88.43 ± 7.87	–	–
	Overall Score	87.71 ± 5.59	–	–
FD_Jenkinson		0.06 ± 0.05	0.09 ± 0.06	0.065^{t}

SaO_2 < 90%, percentage of total sleep time spent at an oxygen saturation < 90%; $LSaO_2$, lowest oxygen saturation; AHI, apnea–hypopnea index; OAI, obstructive apnea index; HI, hypopnea index; AI, arousal index; BMI, body mass index; WISC, the Wechsler Intelligence Scale for Children; VIQ, verbal intelligence quotient; PIQ, performance intelligence quotient; SAFE, Social Adaptive Functioning Evaluation Scale; DQ, developmental quotient; GMQ, gross motor quotient; FMQ, fine motor quotient; t, two-sample t-test; χ^2, chi-square test. Data are presented as mean ± standard deviation.

were right-handed. The exclusion criteria included (1) history of brain structural injury, neurological or psychiatric disorders; (2) suffering from cardiovascular diseases, neuromuscular diseases, or defined genetic syndromes; (3) abnormal blood pressure, blood fat, and glucose; (4) being with any known acute or chronic illness; and (5) undergone treatment with drugs and surgery. Before the scan, children under 7 years old needed to take chloral hydrate for sedation. The dosage was 0.5 ml/kg and the maximum dose was 10 ml. This study was approved by the Medical Ethics Committee of Beijing Children's Hospital, Beijing, China. The study was carried out in line with relevant guidelines by the Medical Ethics Committee of Beijing Children's Hospital, which include MRI scan and clinical diagnosis.

Polysomnography

All subjects underwent a polysomnography (PSG) evaluation (Compumedics E; Compumedics, Melbourne, Australia; or ALICE 5; Philips Respironics, Amsterdam, Netherlands), which recorded a polysomnogram of more than 7.5 h. Simultaneous monitoring was included for EEG, bilateral electro-oculogram, electromyogram of mentalis activity and bilateral anterior tibialis, ECG, arterial oxyhemoglobin saturation and plethysmographic signal by pulse oximetry, heat-sensitive airflow and nasal pressure, chest and abdominal movements, snoring sensor, body position, and other indicators. All sleep monitoring results were scored manually by experienced professional pediatric PSG technicians according to the diagnostic criteria published by the American Academy of Sleep Medicine (AASM) (Berry et al., 2012). The criteria for the OSA group diagnosis consisted of obstructive apnea (OAI) > 1 times/h or apnea–hypopnea index (AHI) > 5 times/h and lowest oxygen saturation ($LSaO_2$) < 92% by PSG.

Image Acquisition

Magnetic resonance imaging scanning was performed on a 3-T MR scanner (GE Medical Systems, Discovery MR750). Before the scan, all participants should keep respiration and heart rate in a normal state. All participants were required to be awake and quietly breathing until the end of the scan. The scanner parameters for fMRI data are TR/TE = 2000/24 ms, 240 time points, image matrix = 64 × 64, voxel size = 3.5 mm × 3.5 mm × 3.5 mm, field of view (FOV) = 224 mm × 224 mm. The scanner parameters for T1-weighted images are TR/TE = 8.19/3.78 ms, voxel size = 0.4688 mm × 0.4688 mm × 1 mm, matrix = 512 × 512, FOV = 240 mm × 240 mm. During MRI scanning, participants were asked to close their eyes and lie still in the scanner. Head positioning was standardized using canthomeatal

landmarks. The head was stabilized with foam pads to minimize its movement.

Data Preprocessing

Resting-state fMRI data reprocessing was performed using the statistical parametric mapping (SPM8) and Data Processing & Analysis for Resting-state Brain Imaging (DPABI Version 2.1)[1]. The first 10 image volumes of functional images were removed for the signal equilibrium and subject's adaptation to the scanning noise. Then, the functional images were corrected for time offsets between slices and geometrical displacements due to head motion. We further calculated the mean frame-wise displacement (FD) to measure voxel-wise differences in motion in its derivation (Jenkinson et al., 2002). None of the participants were excluded based on the excluding criteria of 3.0 mm and 3.0 degree in max head motion, with mean FD > 0.2 mm. The T1-weighted images were co-registered to the average functional images and then segmented into the white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) by using the New Segment tool in DPABI. We removed linear trends and regressed out several nuisance signals from each voxel's time course, including 24-parameter head-motion profiles (Friston et al., 1996; Yan et al., 2013), mean WM, and cerebrospinal fluid (CSF) time series within the respective brain masks derived from prior probability maps in SPM8 (threshold = 0.8). All the corrected functional data were then normalized by *DARTEL* (Ashburner, 2007) to the Montreal Neurological Institute (MNI) space using an optimum 12-parameter affine transformation and non-linear deformations and then resampled to a 3-mm isotropic resolution.

Measurement of Amplitude of Low-Frequency Fluctuation/Fractional ALFF and Regional Homogeneity

To calculate ALFF, we firstly performed the spatial smoothing on the resampled images with a 4-mm full width at half maximum (FWHM) Gaussian kernel. We then converted the smoothed signal of each voxel from time domain to frequency domain via Fast Fourier Transform (FFT) to obtain the power spectrum. This power spectrum (frequency range: 0–0.25 Hz) was square-rooted at each frequency, and then averaged across 0.01–0.08 Hz at each voxel, which was taken as ALFF (Zang et al., 2007). To calculate fALFF, the sum of the amplitude (square root of power spectrum) across 0.01–0.08 Hz was divided by that of the entire frequency range (0–0.25 Hz) (Zou et al., 2008). ALFF/fALFF of each voxel was divided by the global mean ALFF/fALFF for standardization purpose, and mALFF/mfALFF was obtained as a parameter for further statistical comparison and analysis.

Regional homogeneity maps were generated before spatial smoothing. After normalization, the band-pass filtering (0.01–0.08 Hz) was performed on the normalized images to reduce the effects of low-frequency drift and high-frequency physiological noise. ReHo maps were conducted by calculating the Kendall coefficient of concordance (KCC) as synchronization of fMRI signals of nearest neighboring 27 voxels (Zang et al., 2004). For standardization purposes, the ReHo value of each voxel was divided by the whole brain mean ReHo value, and then smoothing was done with a 4 mm FWHM Gaussian kernel. The smReHo map was obtained as the ReHo parameter for further statistical comparison and analysis.

Receiver Operating Characteristic Curves Analysis

Once significantly altered ALFF/fALFF/ReHo areas were found between groups, they might be utilized as markers to discriminate OSA from controls, as useful diagnostic neuroimaging biomarkers (Li et al., 2015). To test this possibility, the mean ALFF/fALFF/ReHo values of significantly altered brain clusters were extracted and used for analysis of the receiver operating characteristic (ROC) curves, using the MedCalc Statistical Software[2]. To summarize the overall diagnostic ability of the tests, we computed the maximum Youden index (sensitivity + specificity − 1) (Fluss et al., 2005), and corresponding sensitivity, specificity, and 95% confidence intervals (CIs) for each cluster.

Statistical Analysis

Further statistical analysis was performed based on a 90% group mask (meaning 90% of subjects have this voxel) generated in *DPABI* toolbox to detect group differences. We conducted a two-sample *t*-test to compare whole-brain ALFF, ALFF and ReHo values between OSA children and controls, including age, gender, and mean FD (Jenkinson et al., 2002) as covariates. The Gaussian random field (GRF) correction (Bansal and Peterson, 2018) was used to correct for multiple comparisons, and the statistical threshold was set at a voxel level of $p < 0.001$ with a cluster-level $p < 0.05$ (two-tailed) in *DPABI* toolbox. All coordinates were reported in MNI space. Brain regions with significant intergroup differences in ALFF/fALFF/ReHo were defined as regions of interest (ROIs). We extracted the mean ALFF/fALFF/ReHo values of these ROIs from OSA children. Partial correlations controlling for age, gender, and FD values were applied in SPSS version 24.0 (SPSS Inc., Chicago, IL, United States) to identify mean ALFF/fALFF/ReHo values related to clinical parameters of OSAS children, using the statistical threshold of $p < 0.05$.

RESULTS

Demographic and Clinical Characteristic

Demographic and clinical characteristics of each group are summarized in **Table 1**. The OSA had significantly higher scores for AHI, OAI, HI, SaO_2 < 90%, and AI and significantly lower scores for $LSaO_2$ and sleep efficiency than controls. No significant differences were found in age, gender, weight, BMI, and mean FD between the two groups. Of note, there is no significant Pearson correlation between mean FD and age for the OSA group ($r = 0.197$, $p = 0.289$) and control group ($r = 0.204$, $p = 0.255$).

[1]http://www.restfmri.net

[2]https://www.medcalc.org

Amplitude of Low-Frequency Fluctuation Results

Compared with controls, the OSA group showed significant lower ALFF in the cluster of the right lingual gyrus (Brodmann area 18). The details are presented in **Table 2** and **Figure 1**.

Fractional ALFF Results

Compared with controls, the OSA group showed significant lower fALFF in the cluster of the left middle frontal gyrus and higher fALFF in the left precuneus. The details are presented in **Table 3** and **Figure 2**.

Regional Homogeneity Results

Compared with controls, OSA group showed significant lower ReHo in the cluster of the left inferior frontal gyrus (orbital part) and left middle frontal gyrus. The details are presented in **Table 4** and **Figure 3**.

Receiver Operating Characteristic Curves Analysis Results

In the ROC curve analysis, the mean fALFF value of left middle frontal gyrus achieves the highest sensitivity, specificity, and area under the ROC curve (AUC). All the altered brain clusters achieve the significance level $p < 0.001$ of AUC, indicating these findings as potential useful diagnostic biomarkers. The details are presented in **Table 5** and **Figure 4**.

Correlation Results

In the OSA group, the mean fALFF value in the left middle frontal gyrus showed significantly negative correlations with HI ($r = -0.385$, $p = 0.043$), $SaO_2 < 90\%$

TABLE 2 | Two-sample *t*-tests demonstrated regions with significantly decreased ALFF in OSA children compared with controls (with GRF correction, voxel level $p < 0.001$, cluster level $p < 0.05$).

Condition	Brain regions	Cluster size	*t*-score of peak voxel	MNI coordinates of peak voxel		
				x	y	z
OSA < HC	Right lingual gyrus (BA 18)	64	5.25	6	−72	−9

HC, healthy children; BA, Brodmann area.

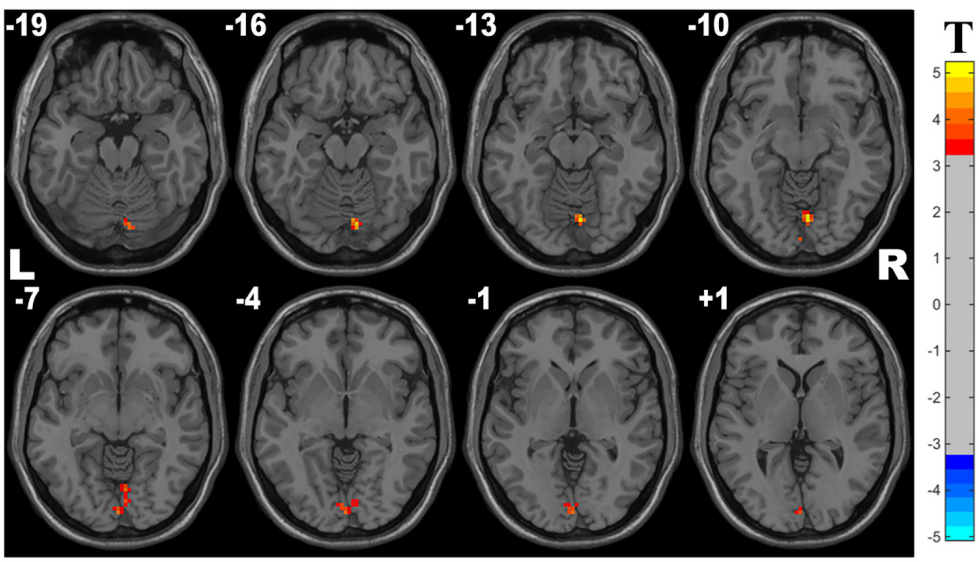


FIGURE 1 | Cluster with significantly decreased ALFF in OSA children compared with controls (with GRF correction, voxel level $p < 0.001$, cluster level $p < 0.05$).

TABLE 3 | Two-sample *t*-tests demonstrated regions with significantly altered fALFF in OSA children compared with controls (with GRF correction, voxel level $p < 0.001$, cluster level $p < 0.05$).

Condition	Brain regions	Cluster size	*t*-score of peak voxel	MNI coordinates of peak voxel		
				x	y	z
OSA > HC	Left precuneus	32	3.98	−18	−75	18
OSA < HC	Left middle frontal gyrus	22	4.34	−27	0	45

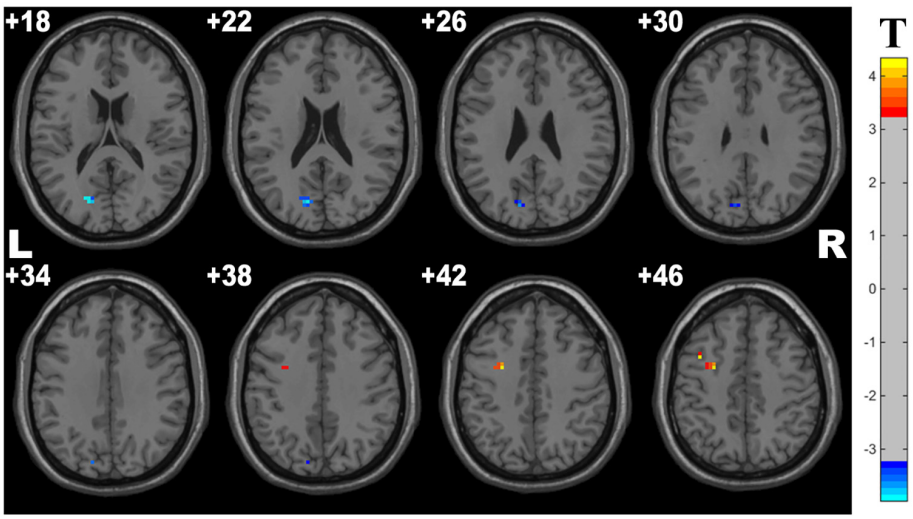


FIGURE 2 | Cluster with significantly altered fALFF in OSA children compared with controls (with GRF correction, voxel level $p < 0.001$, cluster level $p < 0.05$). The blue and red areas denote higher and lower fALFF in OSA group.

TABLE 4 | Two-sample *t*-tests demonstrated regions with significantly decreased ReHo in OSA children compared with controls (with GRF correction, voxel level $p < 0.001$, cluster level $p < 0.05$).

Condition	Brain regions	Cluster size	*t*-score of peak voxel	MNI coordinates of peak voxel		
				x	y	z
OSA < HC	Left inferior frontal gyrus, orbital part	50	4.15	−45	48	−6
OSA < HC	Left middle frontal gyrus	58	4.39	−45	33	27

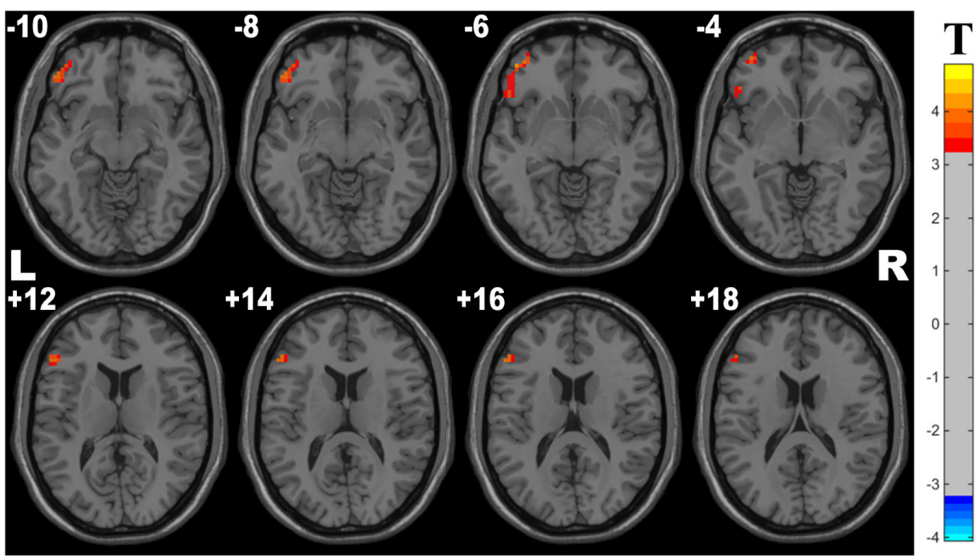


FIGURE 3 | Cluster with significantly decreased ReHo in OSA children compared with controls (with GRF correction, voxel level $p < 0.001$, cluster level $p < 0.05$).

($r = -0.381$, $p = 0.046$), AI ($r = -0.391$, $p = 0.040$), and a significantly positive correlation with $LSaO_2$ ($r = 0.378$, $p = 0.047$). The mean fALFF value in the left precuneus showed a significantly negative correlation with $LSaO_2$ ($r = -0.506$, $p = 0.006$) and a significantly positive correlation with $SaO_2 < 90\%$ ($r = 0.396$, $p = 0.037$) (shown in **Figure 5**). All the p-values were uncorrected for multiple comparisons.

TABLE 5 | The statistics of ROC curve analysis for altered brain clusters.

Clusters	SEN	SPE	AUC	95% CIs
ALFF_Right lingual gyrus	83.87%	78.79%	0.822	0.706–0.906
fALFF_Left precuneus	70.97%	69.70%	0.732	0.607–0.835
fALFF_Left middle frontal gyrus	83.87%	84.85%	0.871	0.764–0.942
ReHo_Left ORBinf	74.19%	84.85%	0.790	0.670–0.882
ReHo_Left middle frontal gyrus	70.97%	78.79%	0.802	0.683–0.891

SEN/SPE, sensitivity/specificity corresponding to maximum Youden index; AUC, area under the ROC curve; CIs, confidence intervals; ORBinf, inferior frontal gyrus, orbital part.

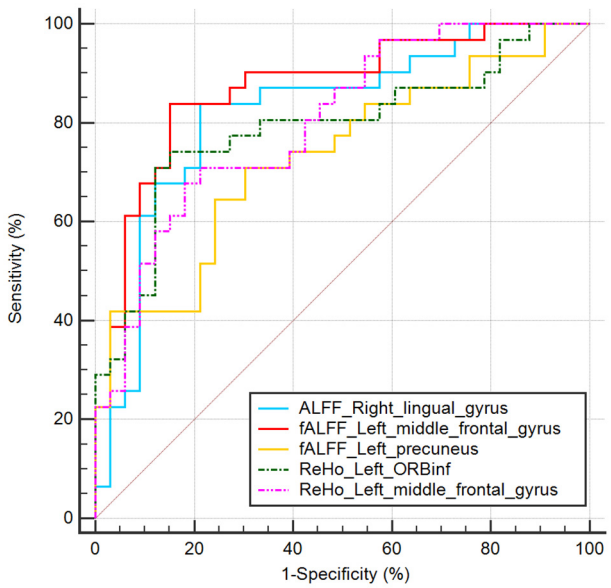


FIGURE 4 | ROC curve analysis of mean ALFF/fALFF/ReHo signal values for altered brain clusters. AUC, area under the curve; ORBinf, inferior frontal gyrus, orbital part.

DISCUSSION

Amplitude of Low-Frequency Fluctuation/Fractional ALFF and ReHo approaches are used to investigate the intrinsic brain activities in OSA, as effective noninvasive imaging tools. These methods are based on different neurophysiological mechanisms, ALFF/fALFF analysis represents neural intensity, whereas ReHo reflects neural coherence (Liu et al., 2017). Our study is the first to utilize the combination of three methods to detect abnormal neural activities in certain brain regions in pediatric OSA, which contributed to the exhibition of more comprehensive functional alterations. Compared to previous study (Li et al., 2015), we applied more clinical parameters analysis and obtained more outstanding correlations with functional parameters. Moreover, the ROC curve analysis was demonstrated in detail, along with sensitivity and specificity, which provided remarkable guidance for clinical diagnosis. In addition, a large sample size was enrolled with 31 OSA individuals in this study. Taken together, we believe our study contributes to obtain more reliable results for revealing the abnormalities of neural activity in OSA children.

Compared with healthy controls, the OSA group showed significant lower ALFF in the right lingual gyrus (Brodmann area 18). In the OSA group, fALFF was significantly decreased in the left middle frontal gyrus as well as increased in the left precuneus. We also observed that the OSA group showed a significant decrease of ReHo in the left inferior frontal gyrus (orbital part) and left middle frontal gyrus. These results revealed the changes of spontaneous brain activities were closely associated with the DMN, the frontal lobe, and the right lingual gyrus. Furthermore, in the OSA group, the mean fALFF value in the left middle frontal gyrus showed significantly negative correlations with HI, SaO_2 < 90%, and AI and a significantly positive correlation with $LSaO_2$. The mean fALFF value in the left precuneus showed a significantly negative correlation with $LSaO_2$ and a significantly positive correlation with SaO_2 < 90%. Additionally, in the previous studies (Zang et al., 2007; Zou et al., 2008), scholars have found the ALFF method may be affected by physiological noise irrelevant to brain activity. However, the artifacts from non-specific brain areas (the ventricles and the vicinity of blood vessels) were significantly reduced by the fALFF approach, while signals from cortical regions correlated with brain activity were enhanced, which contributed to the superior sensitivity and specificity in detecting spontaneous brain activities (Zou et al., 2008). In line with previous researches, in our study, we also noted the fALFF approach showed more rewarding results among these approaches. The mean fALFF value not only significantly associated with the sleep parameters but also demonstrated the best performance in ROC curve analysis.

Our study revealed the increase of fALFF in the left precuneus. The precuneus is located in the posteromedial cortex of the parietal lobe and plays a key role in a range of highly integrated tasks, including visuospatial imagery, self-processing operations, and episodic memory retrieval, namely, first-person perspective taking and an experience of agency (Cavanna and Trimble, 2006). Peng et al. (2014) found that patients with OSA showed decreased ReHo in the precuneus, besides the significant negative correlation between altered ReHo in the precuneus with sleep time, suggesting decreased sleep time might be a key factor for dysfunction in the precuneus. An 18F-fluoro-2-deoxy-D-Glucose positron emission tomography study showed a decrease of brain metabolism in the precuneus (Yaouhi et al., 2009). Another functional imaging study displayed lower ALFF in the right precuneus; in addition, a significant positive correlation was found in the right precuneus between MoCA score and the ALFF value, which indicated the abnormalities of the precuneus may be associated with a cognitive dysfunction (Li et al., 2015).

As is well known, the precuneus is a vital node in the default mode networks (DMN), and others include the posterior cingulate cortex; medial prefrontal cortex; and medial, lateral, and inferior parietal regions. These brain regions are highly interconnected to form DMN, a large-scale network, which is thought to be involved in an array of advanced cognitive functions, such as visuospatial imagery, consciousness, attention, adaptation, episodic memory, executive cognitive control, and behavioral inhibition (Andrews-Hanna et al., 2010; Wilson et al., 2010). As similarly reported in the past, abnormal inactivation in the DMN among patients with OSA was found

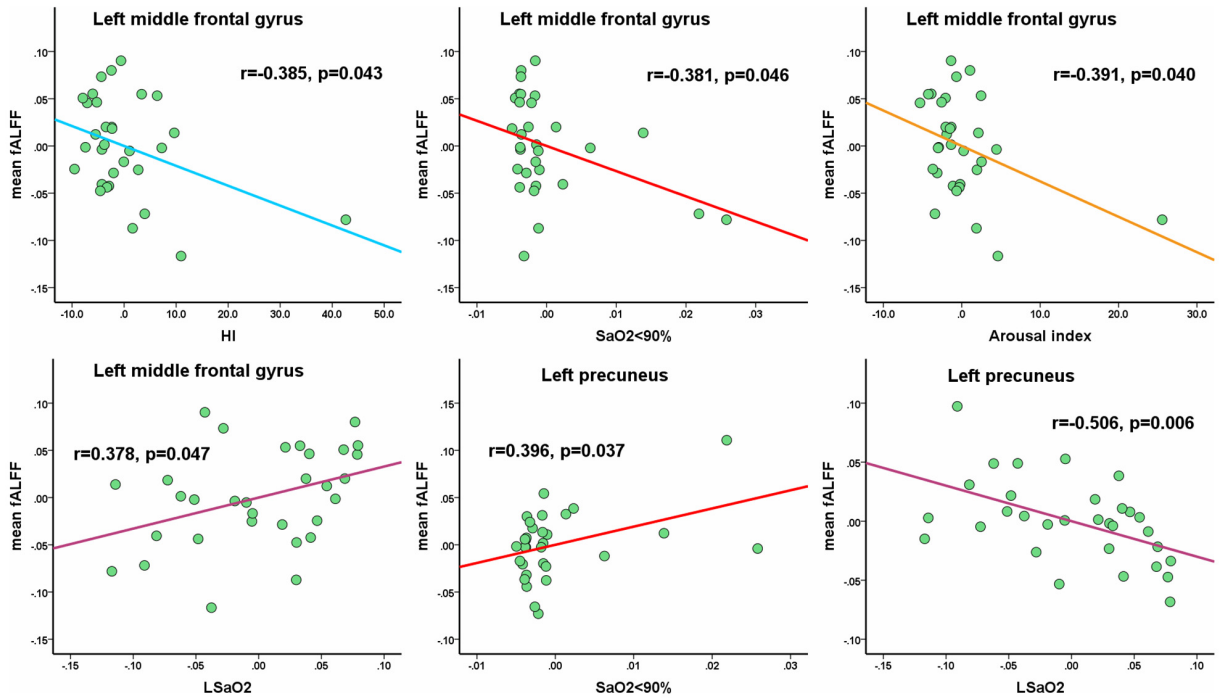


FIGURE 5 | The mean fALFF values in altered clusters show the significantly partial correlations with clinical parameters. Of note, the coordinate value of both *X*-axis (clinical parameter) and *Y*-axis (fALFF value) do not reflect the initial values of these variables, while considering age, gender and FD as covariates. All the *p*-values were uncorrected for multiple comparisons.

during working memory tasks and also significantly positively correlated with behavioral performance, which may imply that the inhibition of activity in the DMN plays a role in cognitive impairment (Prilipko et al., 2011). Specifically, reduced functional connectivity is related to cognitive dysfunction in the DMN at rest and the enhanced functional connectivity of the OSA may be a compensatory mechanism for the decrease of cognition (Chen et al., 2016). Chen et al. (2018) found that the OSA group showed significantly decreased FC of the anterior–posterior DMN and within the posterior DMN and was associated with the MoCA score, using graph theory approaches. Furthermore, they found abnormal FC within the DMN may contribute to the topological reconfiguration of the DMN in patients with OSA, which illuminated the cognitive dysfunction and topological reconfiguration in OSA. Nevertheless, based on these findings, we found that patients with OSA had increased fALFF in the left precuneus, which may imply an adaptive compensatory response in the DMN. Furthermore, our study also found a significant negative correlation between the altered fALFF in the local region of the DMN and the $LSaO_2$, suggesting that intermittent hypoxia may be a principal element for the DMN abnormity in OSA.

Our study showed fALFF and ReHo were decreased in the left middle frontal gyrus and the left inferior frontal gyrus (orbital part), and left middle frontal gyrus, respectively. To our knowledge, the frontal lobe is the higher cortex of cognitive executive function and also plays an important role in emotion, language, attention, working memory, problem solving, impulse control, and social behavior (Stern et al., 2001; Johnson et al., 2009; Nie et al., 2017). Huang et al. (2019) found both decreased gray matter volume (GMV) and functional response in the orbital frontal cortex (OFC) in patients with OSA, which may indicate that OFC is a vulnerable and sensitive area in the brain. Another study demonstrated that patients with OSA showed significantly lower ReHo in the right medial frontal gyrus and right superior frontal gyrus (Peng et al., 2014). Ji et al. (2021) also found significant lower ReHo in the left medial superior frontal gyrus and was positively associated with VIQ, which reflected that such brain area may play a crucial role in the cognitive processing related to VIQ. The inferior frontal gyrus was attributed to the language function, decision making under risk, and the regulation of cognitive control (Foland-Ross and Gotlib, 2012; Luo et al., 2015). Song et al. (2018) found that the impaired FC between the caudate and inferior frontal gyrus may become the basis of the cognitive regulation defects of emotion, which leads to comorbid mood disorders in OSA.

Consistent with these findings, in our study, we found that patients with OSA showed significantly altered brain activities in the left frontal gyrus when compared with controls. Notably, the mean fALFF value of left middle frontal gyrus achieves the highest sensitivity, specificity, and AUC value in ROC curve analysis. Also, significant correlations were found between the mean fALFF values of the left middle frontal gyrus with the more clinical parameters, suggesting the middle frontal gyrus might be a sensitive region in the brain among OSA children, which might turn into the potential useful diagnostic biomarkers.

In the present study, we also observed significant lower ALFF in the right lingual gyrus among patients with OSA. The lingual gyrus of the occipital lobe is located between the calcarine sulcus and the posterior portion of the collateral sulcus and then extends to the tentorial surface of the temporal lobe and joins the hippocampus (Joo et al., 2007). A study by Luo et al. (2015) found that the OSA showed a tendency of decreased degree in the right lingual gyrus, as a topological alteration in and regional properties in patients with OSA. Our study is also consistent with their results of different modalities. Besides, Joo et al. (2007) found that the cerebral blood flow of parahippocampal and lingual gyrus were reduced in OSA during wakefulness, using the 99mTc-ethyl cysteinate dimer (ECD) single photon emission computed tomography method, which partly indicated memory impairment and spatial learning deficits in patients with severe OSA. The lingual gyrus has been believed to be involved in visual recognition and episodic memory consolidation (Kukolja et al., 2016) and are considered to play a role in the process of generating and recalling dreams as well (Bischof and Bassetti, 2004). Therefore, the decreased metrics of lingual gyrus possibly account for certain deficits in visual memory and learning.

LIMITATIONS

Several limitations deserve to be mentioned in the present study. Firstly, our study obtained a larger sample size compared with previous neuroimaging researches in pediatric OSA (Luo et al., 2015; Ji et al., 2021). However, the sample size was still relatively small, which implies an urgent need for expanding data to validate the results of future studies. Secondly, the *p*-values of the correlations between network properties and clinical measurements are not corrected for multiple comparisons in this study, as this is currently a preliminary exploratory research. We will reveal the correlations corrected for multiple comparisons based on more subjects in future research, which obtains the more reliable mechanisms of brain activity in OSA children. Thirdly, the previous study found that the test–retest reliability of PerAF is better than ALFF and fALFF, and the test–retest reliability between machines is better (Zhao et al., 2018). We will investigate this using perAF to measure the spontaneous brain activity of OSA children in our future study.

CONCLUSION

We have investigated the alterations of spontaneous neural activity in OSA, based on the ALFF/fALFF and ReHo approaches on rs-fMRI data. Abnormal regions with the altered neural activity in OSA children include the precuneus, the middle and inferior frontal gyrus, and the lingual gyrus. Moreover, we also found that the altered fALFF in the precuneus was negatively correlated with the $LSaO_2$ in OSA. These results expounded the underlying neurophysiological mechanisms of altered spontaneous brain activity and revealed the correlation between the changes of sleep function and functional activity in brain.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Beijing Children's Hospital, Beijing, China. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JB: study concept and design, investigation, acquisition of data, analysis and interpretation of data, and drafting of the original manuscript. HW: study concept and design, methodology, software, analysis and interpretation of data, and writing – review and editing. JT: collect clinical information and access cognitive function. YP and XN: study supervision and project administration. HL, LM, TJ, and XL: collect clinical information and access cognitive function. YaL: acquisition of data. YuL: study concept and design, and writing – review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (32100902 and 82000991), the Fundamental Research Funds for the Central Universities (SWU118065), and the Beijing Hospitals Authority Youth Programme (code: QML20201206).

ACKNOWLEDGMENTS

The authors would like to thank all the subjects and their families for the time and effort they dedicated to our research and all the experts who helped to complete this research.

REFERENCES

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron* 65, 550–562. doi: 10.1016/j.neuron.2010.02.005

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Bansal, R., and Peterson, B. S. (2018). Cluster-level statistical inference in fMRI datasets: the unexpected behavior of random fields in high dimensions. *Magn. Reson. Imaging* 49, 101–115. doi: 10.1016/j.mri.2018.01.004

Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., et al. (2012). Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. Deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *J. Clin. Sleep Med.* 8, 597–619. doi: 10.5664/jcsm.2172

Bischof, M., and Bassetti, C. L. (2004). Total dream loss: a distinct neuropsychological dysfunction after bilateral PCA stroke. *Ann. Neurol.* 56, 583–586. doi: 10.1002/ana.20246

Cavanna, A. E., and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* 129, 564–583. doi: 10.1093/brain/awl004

Chen, L., Fan, X., Li, H., Ye, C., Yu, H., Gong, H., et al. (2018). Topological reorganization of the default mode network in severe male obstructive sleep apnea. *Front. Neurol.* 9:363. doi: 10.3389/fneur.2018.00363

Chen, T., Yang, M., Liu, B., Liu, Y. T., Zhang, H. X., Liu, C. C., et al. (2016). The resting-state functional connectivity of the default mode networks in patients with obstructive sleep apnea-hypopnea syndrome. *CNS Neurol. Disord. Drug Targets*. doi: 10.2174/1871527316011701112220000

Cheng, J., Yang, H., and Zhang, J. (2019). Donepezil's effects on brain functions of patients with Alzheimer disease: a regional homogeneity study based on resting-state functional magnetic resonance imaging. *Clin. Neuropharmacol.* 42, 42–48. doi: 10.1097/WNF.0000000000000324

Copes, K., and Rosentswieg, J. (1972). The effects of sleep deprivation upon motor performance of ninth-grade students. *J. Sports Med. Phys. Fitness* 12, 47–53.

Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biom. J.* 47, 458–472. doi: 10.1002/bimj.200410135

Foland-Ross, L. C., and Gotlib, I. H. (2012). Cognitive and neural aspects of information processing in major depressive disorder: an integrative perspective. *Front. Psychol.* 3:489. doi: 10.3389/fpsyg.2012.00489

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312

Hamasaki, U. S., Nagata, P. S., Fujita, R. R., Moreira, G. A., Pradella-Hallinan, M., and Weckx, L. (2007). Assessment of cognitive learning function in children with obstructive sleep breathing disorders. *Braz. J. Otorhinolaryngol.* 73, 315–320. doi: 10.1016/S1808-8694(15)30074-4

Huang, X., Tang, S., Lyu, X., Yang, C., and Chen, X. (2019). Structural and functional brain alterations in obstructive sleep apnea: a multimodal meta-analysis. *Sleep Med.* 54, 195–204. doi: 10.1016/j.sleep.2018.09.025

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Ji, T., Li, X., Chen, J., Ren, X., Mei, L., Qiu, Y., et al. (2021). Brain function in children with obstructive sleep apnea: a resting-state fMRI study. *Sleep* 44:zsab047. doi: 10.1093/sleep/zsab047

Johnson, S. B., Blum, R. W., and Giedd, J. N. (2009). Adolescent maturity and the brain: the promise and pitfalls of neuroscience research in adolescent health policy. *J. Adolesc. Health* 45, 216–221. doi: 10.1016/j.jadohealth.2009.05.016

Joo, E. Y., Tae, W. S., Han, S. J., Cho, J. W., and Hong, S. B. (2007). Reduced cerebral blood flow during wakefulness in obstructive sleep apnea-hypopnea syndrome. *Sleep* 30, 1515–1520. doi: 10.1093/sleep/30.11.1515

Kheirandish-Gozal, L., Sahib, A. K., Macey, P. M., Philby, M. F., Gozal, D., and Kumar, R. (2018). Regional brain tissue integrity in pediatric obstructive sleep apnea. *Neurosci. Lett.* 682, 118–123. doi: 10.1016/j.neulet.2018.06.002

Kirk, V., Baughn, J., D'Andrea, L., Friedman, N., Galion, A., Garetz, S., et al. (2017). American academy of sleep medicine position paper for the use of a home sleep apnea test for the diagnosis of OSA in children. *J. Clin. Sleep Med.* 13, 1199–1203. doi: 10.5664/jcsm.6772

Krysta, K., Bratek, A., Zawada, K., and Stepanczak, R. (2017). Cognitive deficits in adults with obstructive sleep apnea compared to children and adolescents. *J. Neural Transm. (Vienna)* 124, 187–201. doi: 10.1007/s00702-015-1501-6

Kukolja, J., Goreci, D. Y., Onur, O. A., Riedl, V., and Fink, G. R. (2016). Resting-state fMRI evidence for early episodic memory consolidation: effects of age. *Neurobiol. Aging* 45, 197–211. doi: 10.1016/j.neurobiolaging.2016.06.004

Li, H. J., Dai, X. J., Gong, H. H., Nie, X., Zhang, W., and Peng, D. C. (2015). Aberrant spontaneous low-frequency brain activity in male patients with severe obstructive sleep apnea revealed by resting-state functional MRI. *Neuropsychiatr. Dis. Treat.* 11, 207–214. doi: 10.2147/NDT.S73730

Li, H. J., Nie, X., Gong, H. H., Zhang, W., Nie, S., and Peng, D. C. (2016). Abnormal resting-state functional connectivity within the default mode network subregions in male patients with obstructive sleep apnea. *Neuropsychiatr. Dis. Treat.* 12, 203–212. doi: 10.2147/NDT.S97449

Liu, Y., Wang, J., Zhang, J., Wen, H., Zhang, Y., Kang, H., et al. (2017). Altered spontaneous brain activity in children with early tourette syndrome: a resting-state fMRI Study. *Sci. Rep.* 7:4808. doi: 10.1038/s41598-017-04148-z

Luo, Y. G., Wang, D., Liu, K., Weng, J., Guan, Y., Chan, K. C., et al. (2015). Brain structure network analysis in patients with obstructive sleep apnea. *PLoS One* 10:e139055. doi: 10.1371/journal.pone.0139055

Macey, P. M., Kheirandish-Gozal, L., Prasad, J. P., Ma, R. A., Kumar, R., Philby, M. F., et al. (2018). Altered regional brain cortical thickness in pediatric obstructive sleep apnea. *Front. Neurol.* 9:4. doi: 10.3389/fneur.2018.00004

Nie, S., Peng, D. C., Gong, H. H., Li, H. J., Chen, L. T., and Ye, C. L. (2017). Resting cerebral blood flow alteration in severe obstructive sleep apnoea: an arterial spin labelling perfusion fMRI study. *Sleep Breath.* 21, 487–495. doi: 10.1007/s11325-017-1474-9

Peng, D. C., Dai, X. J., Gong, H. H., Li, H. J., Nie, X., and Zhang, W. (2014). Altered intrinsic regional brain activity in male patients with severe obstructive sleep apnea: a resting-state functional magnetic resonance imaging study. *Neuropsychiatr. Dis. Treat.* 10, 1819–1826. doi: 10.2147/NDT.S67805

Philby, M. F., Macey, P. M., Ma, R. A., Kumar, R., Gozal, D., and Kheirandish-Gozal, L. (2017). Reduced regional grey matter volumes in pediatric obstructive sleep apnea. *Sci. Rep.* 7:44566. doi: 10.1038/srep44566

Prilipko, O., Huynh, N., Schwartz, S., Tantrakul, V., Kim, J. H., Peralta, A. R., et al. (2011). Task positive and default mode networks during a parametric working memory task in obstructive sleep apnea patients and healthy controls. *Sleep* 34, 293–301. doi: 10.1093/sleep/34.3.293

Qiu, H., Li, X., Luo, Q., Li, Y., Zhou, X., Cao, H., et al. (2019). Alterations in patients with major depressive disorder before and after electroconvulsive therapy measured by fractional amplitude of low-frequency fluctuations (fALFF). *J. Affect. Disord.* 244, 92–99. doi: 10.1016/j.jad.2018.10.099

Song, X., Roy, B., Kang, D. W., Aysola, R. S., Macey, P. M., Woo, M. A., et al. (2018). Altered resting-state hippocampal and caudate functional networks in patients with obstructive sleep apnea. *Brain Behav.* 8:e994. doi: 10.1002/brb3.994

Stern, C. E., Sherman, S. J., Kirchhoff, B. A., and Hasselmo, M. E. (2001). Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. *Hippocampus* 11, 337–346. doi: 10.1002/hipo.1048

Stevens, D., Jackson, B., Carberry, J., McLoughlin, J., Barr, C., Mukherjee, S., et al. (2020). The impact of obstructive sleep apnoea on balance, gait and falls risk: a narrative review of the literature. *J. Gerontol. A Biol. Sci. Med. Sci.* 75, 2450–2460. doi: 10.1093/gerona/glaa014

Wilson, C. R., Gaffan, D., Browning, P. G., and Baxter, M. G. (2010). Functional localization within the prefrontal cortex: missing the forest for the trees? *Trends Neurosci.* 33, 533–540. doi: 10.1016/j.tins.2010.08.001

Xia, W., Chen, Y. C., Luo, Y., Zhang, D. F., Chen, H., Ma, J., et al. (2018). Decreased spontaneous brain activity and functional connectivity in type 1 diabetic patients without microvascular complications. *Cell. Physiol. Biochem.* 51, 2694–2703. doi: 10.1159/000495960

Yan, C. G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., et al. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *Neuroimage* 76, 183–201. doi: 10.1016/j.neuroimage.2013.03.004

Yaouhi, K., Bertran, F., Clochon, P., Mezenge, F., Denise, P., Foret, J., et al. (2009). A combined neuropsychological and brain imaging study of obstructive sleep apnea. *J. Sleep Res.* 18, 36–48. doi: 10.1111/j.1365-2869.2008.00705.x

Yu, H., Chen, L., Li, H., Xin, H., Zhang, J., Wei, Z., et al. (2019). Abnormal resting-state functional connectivity of amygdala subregions in patients with obstructive sleep apnea. *Neuropsychiatr. Dis. Treat.* 15, 977–987. doi: 10.2147/NDT.S191441

Yu, Y., Li, Z., Lin, Y., Yu, J., Peng, G., Zhang, K., et al. (2019). Depression affects intrinsic brain activity in patients with mild cognitive impairment. *Front. Neurosci.* 13:1333. doi: 10.3389/fnins.2019.01333

Zang, Y., Jiang, T., Lu, Y., He, Y., and Tian, L. (2004). Regional homogeneity approach to fMRI data analysis. *Neuroimage* 22, 394–400. doi: 10.1016/j.neuroimage.2003.12.030

Zang, Y. F., He, Y., Zhu, C. Z., Cao, Q. J., Sui, M. Q., Liang, M., et al. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain Dev.* 29, 83–91. doi: 10.1016/j.braindev.2006.07.002

Zhang, Q., Wang, D., Qin, W., Li, Q., Chen, B., Zhang, Y., et al. (2013). Altered resting-state brain activity in obstructive sleep apnea. *Sleep* 36, 651–659. doi: 10.5665/sleep.2620

Zhang, Z., Liu, Y., Jiang, T., Zhou, B., An, N., Dai, H., et al. (2012). Altered spontaneous activity in Alzheimer's disease and mild cognitive impairment revealed by Regional Homogeneity. *Neuroimage* 59, 1429–1440. doi: 10.1016/j.neuroimage.2011.08.049

Zhao, N., Yuan, L. X., Jia, X. Z., Zhou, X. F., Deng, X. P., He, H. J., et al. (2018). Intra- and inter-scanner reliability of voxel-wise whole-brain analytic metrics for resting state fMRI. *Front. Neuroinform.* 12:54. doi: 10.3389/fninf.2018.00054

Zou, Q. H., Zhu, C. Z., Yang, Y., Zuo, X. N., Long, X. Y., Cao, Q. J., et al. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. *J. Neurosci. Methods* 172, 137–141. doi: 10.1016/j.jneumeth.2008.04.012

Zuo, X. N., Di Martino, A., Kelly, C., Shehzad, Z. E., Gee, D. G., Klein, D. F., et al. (2010). The oscillating brain: complex and reliable. *Neuroimage* 49, 1432–1445. doi: 10.1016/j.neuroimage.2009.09.037

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership