medRxiv preprint doi: https://doi.org/10.1101/2023.04.22.23288741; this version posted April 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Transfer learning on structural brain age models to decode cognition in MS: a federated learning approach

Stijn Denissen[1,2,3,✉], **Matthias Grothe**[4], **Manuela Vaněčková**[3], **Tomáš Uher**[5], **Jorne Laton**[1], **Matěj Kudrna**[3], **Dana Horáková**[5], **Michael Kirsch**[6], **Jiří Motýl**[5], **Maarten De Vos**[7], **Oliver Y. Chén**[8,9], **Jeroen Van Schependom**[1,10], **Diana Maria Sima**[1,2], **and Guy Nagels**[1,2,11]

[1]AIMS Lab, Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium
[2]icometrix, Leuven, Belgium
[3]Department of Radiology, First Faculty of Medicine, Charles University, General University Hospital, Prague, Czech Republic
[4]Department of Neurology, University Medicine Greifswald, Greifswald, Germany
[5]Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University, General University Hospital, Prague, Czech Republic
[6]Institute for Diagnostic Radiology and Neuroradiology, University Medicine of Greifswald, Greifswald, Germany
[7]Departments of Electrical Engineering (ESAT) and Development & Regeneration, KU Leuven, Leuven, Belgium
[8]Département Médecine de Laboratoire et Pathologie (DMLP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[9]Faculté de Biologie et de Médecine (FBM), Université de Lausanne, Lausanne, Switzerland
[10]Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussels, Belgium
[11]St Edmund Hall, University of Oxford, Queen's Lane, Oxford, UK

Introduction: Classical deep learning research requires lots of centralised data. However, data sets are often stored at different clinical centers, and sharing sensitive patient data such as brain images is difficult. In this manuscript, we investigated the feasibility of federated learning, sending models to the data instead of the other way round, for research on brain magnetic resonant images of people with multiple sclerosis (MS).

Methods: Using transfer learning on a previously published brain age model, we trained a model to decode performance on the symbol digit modalities test (SDMT) of patients with MS from structural T1 weighted MRI. Three international centers in Brussels, Greifswald and Prague participated in the project. In Brussels, one computer served as the server coordinating the FL project, while the other served as client for model training on local data (n=97). The other two clients were Greifswald (n=104) and Prague (n=100). Each FL round, the server sent a global model to the clients, where its fully connected layer was updated on the local data. After collecting the local models, the server applied a weighted average of two randomly picked clients, yielding a new global model.

Results: After 22 federated learning rounds, the average validation loss across clients reached a minimum. The model appeared to have learned to assign SDMT values close to the mean with a mean absolute error of 9.04, 10.59 and 10.71 points between true and predicted SDMT on the test data sets of Brussels, Greifswald and Prague respectively. The overall test MAE across all clients was 10.13 points.

Conclusion: Federated learning is feasible for machine learning research on brain MRI of persons with MS, setting the stage for larger transfer learning studies to investigate the utility of brain age latent representations in cognitive decoding tasks.

Multiple Sclerosis | MRI | Federated Learning | Transfer Learning | Brain Age | Cognition

Correspondence: ***stijn.denissen@vub.be***

Introduction

Magnetic resonance imaging (MRI) changed the way medicine is practised. For neurological disorders, MRI is for example useful to obtain anatomical representations of the brain, based on tissue properties such as the time it takes for protons to align back to a magnetic field after being distorted by a radio-frequency pulse. For multiple sclerosis (MS), this allows optimal MS care in terms of diagnosis and follow-up (1, 2), and can already be considered indispensable less than 50 years after Peter Mansfield successfully scanned the finger of his assistant Andrew A. Maudsley (3).

To make sense of the wealth of information that is within these anatomical brain representations, we can extract features that are relevant for a certain pathology, thus creating a new representation. In MS for example, representations related to brain atrophy are relevant, as they are key for disease monitoring (4). Yet, these knowledge-based, structural representations fall short in explaining real-life symptoms that persons with MS experience, which is known as the "clinico-radiological paradox" (5). It is plausible that such representations should be enriched with other biological information, such as functional brain organisation (6). However, besides resolving to other methodologies, recent evidence suggests that more information can be extracted from structural MRI than common knowledge-based representations (7).

Leonardsen et al. 2022 recently showed that we can in fact obtain clinically relevant representations of structural MR images by using the "brain age" concept (7). The authors showed that the latent space representation of a deep convolutional neural network (CNN) predicting age from structural MRI is useful for distinguishing people with MS from healthy controls. In contrast to a knowledge-based representation, this latent space is a data-driven representation, which is typically not interpretable for humans. Although it is unclear whether these representations are a useful alternative to overcome the aforementioned paradox, we recently showed that brain age is related to disease burden of persons with MS in terms of information processing speed, independently of their chronological age (8). Analogously to Leonardsen et al. 2022 (7), we will now use transfer

DRAFT

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.22.23288741; this version posted April 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

learning (adapting a model performing a certain task to perform a related task) to investigate whether the latent space of brain age models could be useful for decoding cognitive performance from structural MRI in MS.

To investigate this, we need to be able to access a sufficiently large data set. However, sharing medical data is difficult because of e.g. privacy issues, hospital regulations and the General Data Protection Regulation (GDPR). A solution to this is the concept of federated learning (FL) (9), where instead of centralising data, models are trained on distributed data sets by sending models to the data, where they are trained locally. The feasibility of federated learning has already been demonstrated in a medical context, where FL reached a comparable performance in tumour segmentation on MR images compared to conventional centralised learning that requires data sharing (10). The feasibility of federated learning was underlined by a recent study on brain tumour segmentation using a world-wide network of 71 sites (11).

The primary aim of this study is to assess the feasibility of federated learning on decentralised international data of persons with MS, and compare the performance with client-specific model training. Our secondary aim is to provide a benchmark for decoding cognitive performance from T1 weighted MR images using federated learning.

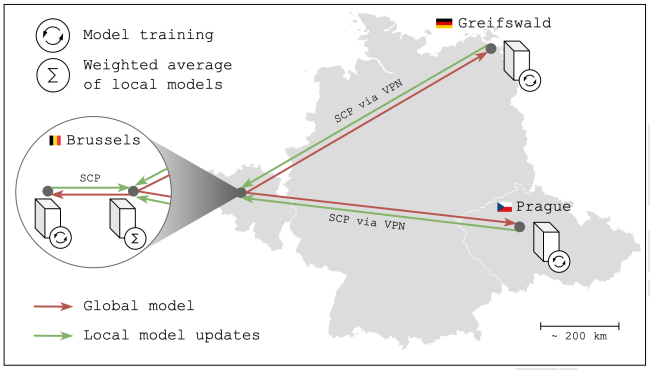

Fig. 1. The federated learning network. The computer with the "sigma" symbol is the server, whereas computers with an "update" symbol are clients. Abbreviations: SCP = Secure Copy Protocol; VPN = Virtual Private Network.

Methods

Study design. This is a cross-sectional study on decentralised data located in Brussels (BE), Greifswald (DE) and Prague (CZ).

Data. This study uses retrospective data collected at each clinical center. For each center in the federated learning network (figure 1), T1 weighted MR images were available, as well as demographic and clinical information. This entailed sex, age, expanded disability status scale (EDSS (12), overall disability), symbol digit modalities test (SDMT (13), explained below), disease duration and MS subtype. Preprocessing of T1 weighted MR images was performed using the pre-processing pipeline of Wood et al. 2022 (14), for which the code was available in their GitHub repository. This pipeline included skull-stripping, registration to Montreal Neurosciences Institute (MNI) 152 space (1mm isotropic) and cropping to a resolution of 130x130x130. The only differences were the use of the Python package "dicom2nifti" v2.3.0 to convert Digital Imaging and Communications in Medicine (DICOM) files to Neuroimaging Informatics Technology Initiative (NIfTI) files, the use of ANTsPyX v0.3.5 since ANTsPyX v0.3.2 was no longer available and the use of a more recent version of PyTorch (15) (v1.13.1) since v1.7.1 did not work with Compute Unified Device Architecture (CUDA) v12.1 (16). Data were organised locally in the Brain Imaging Data Structure (BIDS) format (17). Data are described in table 1.

The SDMT was the target variable to predict. In this test, a subject is presented a list of symbols that need to be converted to numbers using a key on the top of the page, matching symbols with numbers. In 90 seconds, the subject has to convert as many symbols to numbers as possible, each time saying the number out loud for the test administrator to write down. The SDMT is a measure of information processing speed.

	Brussels	Greifswald	Prague	p value
n	97	104	100	
sex (m:f)	28:69	35:69	24:76	0.315†
age (M ± SD)	47.9 ± 9.9	43.1 ± 12.0	44.1 ± 8.6	0.003*
SDMT (M ± SD)	48.1 ± 11.6	51.2 ± 15.0	59.2 ± 10.8	<.001*
EDSS (Median; IQR)	3; 2	1.5; 2	2; 2.125	/
Disease duration (M ± SD)	15.4 ± 8.5	8.4 ± 6.2	14.7 ± 6.5	<.001*
Onset (relapsing:progressive)	90:7	101:3	100:0	0.018†

Table 1. Characteristics of the three different data sets. Abbreviations: n = sample size, m = male, f = female, M = mean, SD = standard deviation, SDMT = symbol digit modalities test, EDSS = expanded disability status scale. P values indicated with a dagger (†) were calculated with a chi-squared test. P values indicated with an asterisk (*) were calculated with an ANalysis Of VAriance (ANOVA) on the sample size (n), mean and standard deviation reported in this table. This method is described in Anders et al. 2016 (18) and was used to avoid data sharing.

Brain age model. We used a pre-trained T1 brain age model from Wood et al. 2022 (14), which the authors made available in their GitHub repository. We chose this model for being a deep neural network that only uses brain images as input, having a low error in predicting age from structural MRI and the ability to replicate their methodology using their code, from data pre-processing to predicting brain age.

Architecture. In brief, the model is a Dense Convolutional Network (DenseNet) (19), which is unique for directly connecting all layers inside the network with each other (19). Although the exact model architecture can be consulted in the paper of Wood et al. 2022 (14), in the context of transfer learning in this manuscript, it is noteworthy to mention the size of the fully connected layer, consisting of 1024 weights and 1 bias (figure 2). These weights were updated during transfer learning, whereas the weights of the deeper layers were frozen.

Age decoding performance. First, we applied the brain age model of Wood et al. 2022 (14) to data of 50 healthy controls from the Brussels client to establish the generalisability

medRxiv preprint doi: https://doi.org/10.1101/2023.04.22.23288741; this version posted April 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

of the model. This was done by calculating the mean absolute error (MAE) between predicted age (brain age) and the chronological age at scanning time. This data set is described in Denissen et al. 2022 (8). Furthermore, as brain age models typically overestimate age of MS patients (8, 20), we also applied the model to the MS data set of each client. For all data sets, we then calculated the brain-predicted age difference (BPAD) by subtracting chronological age from brain age, and tested whether it was significantly different from 0 with a Wilcoxon signed rank test.

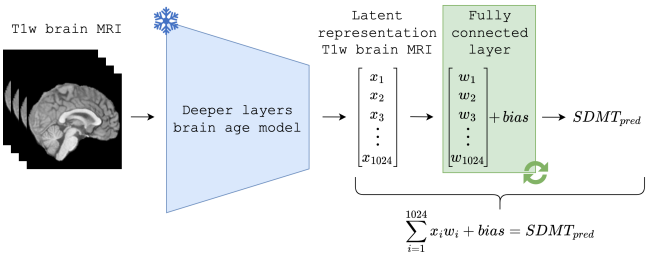

Fig. 2. Transfer learning methodology. The deeper layers of the 3D DenseNet were frozen during training, whereas the fully connected layer (including 1024 weights and 1 bias) was updated. In between the deeper layers and the fully connected layer is the latent (data-driven) representation of a T1w brain MRI.

Hardware setup. The federated learning network (figure 1) consists of 4 computers, of which one is the server that coordinates the project, whereas the other three are clients on which models are trained using the local data that is present. The two Brussels computers were located in the same office and connected to the network of the department of electronics and informatics (ETRO) of the VUB. The computers in Greifswald and Prague were connected to this network via a Virtual Private Network (VPN). Models were shared via secure copy protocol (SCP) with secure shell (SSH). All client computers were equiped with a graphical processing unit (GPU); Brussels: NVIDIA Titan X Pascal (12GB), Greifswald: Zotac RTX GeForce 3090 (24GB) and Prague: INNO3D GeForce RTX 4090 (24GB).

Federated learning. Our federated learning (FL) approach was inspired by the federated averaging (FedAvg) algorithm described in McMahan et al. 2017 (9). Prior to the first federated learning round, the server first sent out a federated learning plan, the latter inspired by the open source OpenFL framework (21). This FL plan contains all details for local model training and can be consulted in the yellow box. Next, each client informed the server about its data set size. The test data for each client was fixed during the entire FL process and only used for testing the final model. Model training happened with FL rounds, where each round consisted of the following steps:

1. The server first sent out a global model to all clients. The initial model was a T1 brain age model (cfr. supra).

2. Next, each client trained the fully connected layer (1024 weights and 1 bias) of the global model using the local, skull-stripped T1 weighted brain MR images as input and SDMT values as ground truth (figure 2), i.e. a regression task. To avoid a lucky split in train and validation data, we used bootstrapping (sampling with replacement) to generate 30 train and validation data sets, yielding 30 models. The model that was sent back to the server was a weighted average of the fully connected layer of these models. Models with a higher validation loss had a lower contribution. Training results (train and validation MAE for every split) were also sent to the server.

3. Lastly, the server randomly sampled 2 local models, and aggregated them using a weighted average, resulting in a new global model for the next FL round. The weight of each local model was determined by the data set size of that client. This concludes the federated learning round.

The best global model across all FL rounds is the one with the lowest average validation MAE across all client models and referred to as the "final model". If a model did not improve for 10 FL rounds, training was stopped early.

Finally, the performance of the final model on unseen data was assessed by applying it to the test data set of each client. Performance was assessed using the MAE and the Pearson correlation between true and predicted SDMT. The overall test MAE was calculated as a weighted average of the test MAE per client:

$$MAEtest_{overall} = \sum_{i=1}^{m} \frac{MAEtest_i * n_i}{N}$$

with m the number of clients, n_i the client sample size and N the summed sample size of all clients.

Client-specific training. As a comparison for the federated learning approach, on each client in our FL network, we performed a client-specific training using only the data set of that client. We used the exact same methodology as for the federated learning approach, but without model averaging across clients. Hence, the client model resulting from each round was immediately passed to the next round. All client models were assessed on the test data set of each client, who shared their test results with the server.

Ethics. The "Commissie Medische Ethiek" (CME) of the UZ Brussel judged this retrospective study to be exempt from ethical approval (B.U.N. 1432022000303). For data at each center in this study, ethical approval was obtained prior to data acquisition (Brussels: B.U.N. 143201423263, Greifswald: BB159/18, Prague: 113/22 S-IV and 28/17), and written informed consent was acquired from all subjects prior to inclusion.

DRAFT

medRxiv preprint doi: https://doi.org/10.1101/2023.04.22.23288741; this version posted April 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

FL plan (Training details)

- **FL rounds**: 100. A federated learning round is one complete cycle of [1] the server sending a global model to all clients, after which [2] all clients update it on their local data and [3] send it back to the server. The FL round is concluded by [4] a weighted average of a certain number of client models (cfr. infra).
- **Number of epochs per FL round**: 1. One epoch is one complete model update on all available training data.
- **Batch size**: 8. Number of data points used simultaneously to calculate the gradient, which allows to update all weights in a model simultaneously.
- **Initial learning rate**: 0.001. The learning rate controls how the model's weights are updated based on the gradient, namely by controlling the magnitude of the step taken into the opposite direction of the gradient.
- **Patience learning rate reduction**: 3. The number of FL rounds without validation loss improvement (tracked per client) before reducing the learning rate by the learning rate reduction factor (cfr. infra).
- **Learning rate reduction factor**: 0.5. Factor by which the learning rate is reduced after several rounds (cfr. supra) without validation loss improvement (tracked per client).
- **Patience early stopping**: 10. Number of FL rounds without improvement of the average validation loss across clients before stopping training early.
- **Train/Validation/Test fraction**: 60/20/20%. Fraction of client data used for the different data sets used for machine learning.
- **Number of clients in sample**: 2. Number of clients of which the local model is used for the weighted average for a new global model.
- **Number of splits**: 30. Number of random train/validation splits (using bootstrapping) for each FL round.
- **Loss function**: L1 loss, $\sum_{i=1}^{n} |y_i - \hat{y}_i|$. Summed absolute error between true and predicted SDMT score.
- **Optimizer**: To update the weights, we used Adam optimisation (22).

Results

Brain age predictions. The brain age model of Wood et al. 2022 (14) achieved an MAE of 3.85 years on the Brussels HC data set, whereon it significantly underestimated age (table 2). The model overestimated age on the Brussels and Greifswald data set (table 2). BPAD distributions of the client MS data sets were significantly different (p <.001, calculated with an ANOVA on n, mean and SD of table 2).

Federated learning. Figure 3 shows the federated learning results. The x-axis shows the number of FL rounds. The y-axis shows the mean absolute error (MAE), which is the L1 loss (sum of absolute differences between true and predicted SDMT value) divided by the sample size. We plotted the MAE instead of the L1 loss since it can be easily interpreted as the "average points of SDMT misprediction by the model". In the top 3 panels, the red and blue lines represent the average train and validation MAE respectively, whereas the shaded red and blue areas represent the 95% confidence interval, all calculated across 30 bootstraps per FL round. It can be observed that both the training and validation MAE are reducing in the first FL rounds, indicating learning behaviour of the model on all three clients. In the bottom panel, the MAE represents the average validation MAE across clients. At FL round 22, this graph reaches a minimum (9.30 points), indicating the best and final model. As training was stopped early after 10 FL rounds of no improvement, the model was trained for a total of 22 + 10 = 32 FL rounds.

	Brussels (HC)	Brussels (MS)	Greifswald (MS)	Prague (MS)
n	50	97	104	100
BPAD (M ± SD)	-2.9 ± 3.7	2.9 ± 8.4	6.1 ± 6.9	0.8 ± 7.0
W (p value)	154 (<.001)	1610 (0.006)	434.5 (<.001)	2238.5 (0.325)

Table 2. Abbreviations: BPAD = brain-predicted age difference, M = mean, SD = standard deviation, W = Wilcoxon signed rank test statistic.

The final model decoded SDMT score with an overall test MAE of 10.13 points, whereas the test MAE per client was 9.04 for Brussels, 10.59 for Greifswald and 10.71 for Prague. The Pearson correlation between true and predicted SDMT was 0.30 (p = 0.206) for Brussels, 0.29 (p = 0.210) for Greifswald and 0.54 (p = 0.014) for Prague.

Client-specific training. Here, we trained a total of three models, one per client. Each model was trained solely on the data that is available locally and tested on all test data sets that were also used for the federated learning approach. The results, expressed as MAE in SDMT points, are displayed in table 3.

		Training data set		
		Brussels	Greifswald	Prague
Test data set	Brussels	**7.68**	**10.57**	12.57
	Greifswald	**9.00**	**9.06**	**9.29**
	Prague	13.61	12.60	**9.00**
	Weighted average	**10.11**	10.72	10.25

Table 3. Client-specific model performance. The columns indicate on which data set a model was trained, while the rows indicate to which test data set a model was applied. Each value is the MAE in SDMT points. Values in bold indicate where the client-specific model training outperformed federated learning.

Discussion

In this manuscript, we showed that federated learning is feasible for training models on T1 weighted brain MR images of people with MS, using an international network of three different clinical centers. The final federated learning

DRAFT

medRxiv preprint doi: https://doi.org/10.1101/2023.04.22.23288741; this version posted April 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

model decoded SDMT score with an overall test MAE of 10.13 points, and a test MAE per client of 9.04 (Brussels), 10.59 (Greifswald) and 10.71 (Prague).

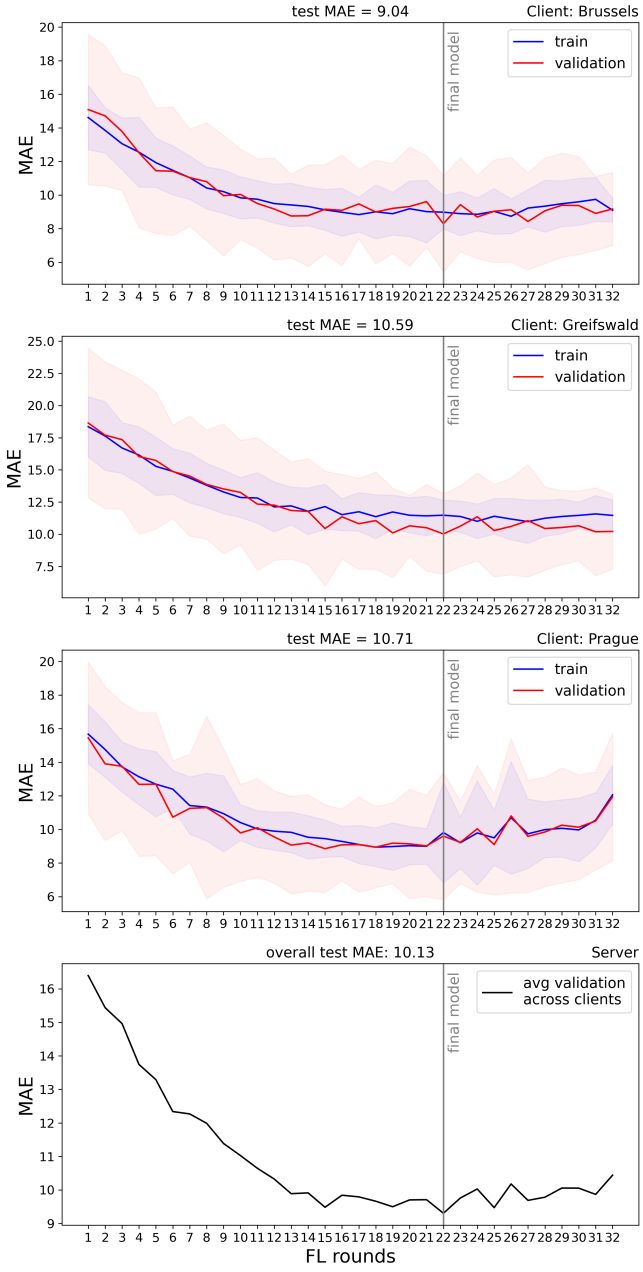

Fig. 3. Federated learning results. Abbreviations: FL = federated learning, MAE = mean absolute error, avg = average.

The federated learning approach. Our approach to federated learning can be considered basic, but its simplicity makes it robust and scalable. Furthermore, our approach is stable, and after setup, only requires starting one Python script per computer involved. However, as we designed our approach to work on a network of 4 computers with Linux installed, we were able to use secure copy protocol (SCP), which only works on UNIX-based operating systems.

Currently, open source federated learning frameworks are in full development, such as Flower (23), OpenFL (21) and PySyft (24). Ultimately, technical developments will increase the number of clients that can be present in a federated learning network. Access to more data sets will in turn allow training deep neural networks on more and heterogeneous data, potentially augmenting generalisability of models. Specifically for constructing cognition decoding models, this will also allow to train deep neural networks from scratch without the need to perform transfer learning on pre-trained networks, such as brain age networks. This might have a couple of advantages, as discussed next.

The performance of the SDMT decoding model. Based on the reduction in MAE over the FL rounds, we observe learning behaviour for each client using the FL approach. Through access to all data sets, federated learning seemed to yield more balanced results, and thus a more generalisable model, compared to a setting where a model only has access to its local data. For this study, a comparison with a centralised learning approach, where all data is centralised, was not possible. Yet, evidence already exists that federated learning on brain MR images yields similar results as centralised learning for tumour segmentation (10).

Besides resolving to federated learning to improve cognition decoding results, a plateauing performance can be expected when starting from brain age models, who are likely insensitive to MS inflammatory activity as these processes are not at play in a healthy aging cohort. Moreover, including inflammatory information is important when decoding information processing speed since besides atrophy, lesion volume is important in the prediction of cognitive impairment (25). As lesions are most clearly visible on FLuid-Attenuated Inversion Recovery (FLAIR) images, these images should be included in future studies decoding cognitive performance from brain MRI.

Although our results appear a fair benchmark when solely considering the MAE, we observed that the Pearson correlation between true and predicted SDMT on the test data set of each client was poor. In a post-hoc analysis, each client therefore shared information on the distributions of the true and predicted SDMT values of the test data set with the server (table 4). The key observation for this table is the standard deviation of the predicted SDMT distribution, which is very low compared to the true SDMT distribution. Hence, the model most probably learned to assign values close to the mean, which yields a fair MAE, but poor individual predictions. When this is indeed the case, we hypothesise this behaviour to be due to the model essentially "giving up" to perform the task with the current resources. The cause could for example be an excess of data heterogeneity, freezing too many network weights (essentially overestimating the similarity of the age and SDMT decoding task), insufficient cognition-related information in a T1 weighted image (cfr. supra) or a lack of data. The latter is exactly why we set up this FL network, which we aim to extend in the future.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.22.23288741; this version posted April 26, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

	Brussels		Greifswald		Prague	
	True	*Pred*	*True*	*Pred*	*True*	*Pred*
Mean	45.6	51.2	49.4	51.4	58.4	51.7
SD	10.3	2.0	13.7	2.0	11.6	2.2
Skewness	-0.24	0.15	-0.12	-0.33	-0.74	-0.21
Kurtosis	-0.19	-0.92	-0.57	0.25	-0.10	-0.88
W	0.94	0.96	0.97	0.96	0.94	0.94
p value	0.290	0.525	0.722	0.567	0.246	0.288

Table 4. Information on distributions of the predicted and true ground truth values of the test data set of each client. Abbreviations: Pred = predicted SDMT, SD = standard deviation, W = Shapiro-Wilk test statistic.

Conclusion

This study showed that federated learning is feasible for machine learning research on MR images in an international network of clinical MS centers, setting the stage for the creation of better models for decoding cognition from MRI in MS while mitigating data sharing.

Funding

This study was funded by a personal industrial PhD grant (Baekeland, HBC.2019.2579) appointed by Flanders Innovation and Entrepreneurship to Stijn Denissen and a personal travel grant (V412023N) appointed by the "Fonds Wetenschappelijk Onderzoek" (FWO) to Stijn Denissen for his stay in Prague in the context of this study. This project was furthermore funded by an IOF-POC grant from the Vrije Universiteit Brussel (VUB), by an ITEA grant (20030 HeKDisco, HBC.2021.0500) from Flanders Innovation and Entrepreneurship, by an institutional support from the Czech Ministry of Health (RVO-VFN 64165) and by the National Institute for Neurological Research, Czech Republic, Programme EXCELES, ID Project No. LX22NPO5107, funded by the European Union – Next Generation EU. Guy Nagels is a senior clinical research fellow of the FWO Flanders (1805620N).

Code availability

In order to support future FL projects, after curation of our code, we plan to make it available in the GitHub repository of our lab, the Artificial Intelligence-supported Modelling in clinical Sciences (AIMS) lab of the Vrije Universiteit Brussel (VUB).

Acknowledgements

We would like to thank Jelle Laton and Robert Malinowski for their help in the setup and maintenance of the hardware present in the federated learning network. This manuscript was created by adapting the Henriques Lab bioRxiv template. Figures 1 and 2 were created using the Mac desktop version of draw.io, and using adapted icons from JGraph, licensed under CC BY 4.0. For figure 1, we also used flags from flag-colorcodes and the Python package GADM v0.0.3.

Bibliography

1. Anthony L Traboulsee and DK Li. The role of MRI in the diagnosis of multiple sclerosis. *Advances in neurology*, 98:125–146, 2006.
2. Ulrike W Kaunzner and Susan A Gauthier. MRI in the assessment and monitoring of multiple sclerosis: an update on best practice. *Therapeutic advances in neurological disorders*, 10 (6):247–261, 2017.
3. Peter Mansfield and Andrew A Maudsley. Medical imaging by NMR. *The British journal of radiology*, 50(591):188–194, 1977.
4. Mike P Wattjes, Àlex Rovira, David Miller, Tarek A Yousry, Maria P Sormani, Nicola De Stefano, Mar Tintore, Cristina Auger, Carmen Tur, Massimo Filippi, et al. Magnims consensus guidelines on the use of MRI in multiple sclerosis–establishing disease prognosis and monitoring patients. *Nature Reviews Neurology*, 11(10):597–607, 2015.
5. Frederik Barkhof. The clinico-radiological paradox in multiple sclerosis revisited. *Current opinion in neurology*, 15(3):239–245, 2002.
6. Martin Sjøgård, Vincent Wens, Jeroen Van Schependom, Lars Costers, Marie D'hooghe, Miguel D'haeseleer, Mark Woolrich, Serge Goldman, Guy Nagels, and Xavier De Tiège. Brain dysconnectivity relates to disability and cognitive impairment in multiple sclerosis. *Human brain mapping*, 42(3):626–643, 2021.
7. Esten H Leonardsen, Han Peng, Tobias Kaufmann, Ingrid Agartz, Ole A Andreassen, Elisabeth Gulowsen Celius, Thomas Espeseth, Hanne F Harbo, Einar A Høgestøl, Ann-Marie de Lange, et al. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage*, 256:119210, 2022.
8. Stijn Denissen, Denis Alexander Engemann, Alexander De Cock, Lars Costers, Johan Baijot, Jorne Laton, Iris-Katharina Penner, Matthias Grothe, Michael Kirsch, Marie Beatrice D'hooghe, et al. Brain age as a surrogate marker for cognitive performance in multiple sclerosis. *European Journal of Neurology*, 29(10):3039–3049, 2022.
9. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
10. Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 92–104. Springer, 2019.
11. Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):1–17, 2022.
12. John F Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1444, 1983.
13. Aaron Smith. *Symbol digit modalities test*. Western psychological services Los Angeles, 1973.
14. David A Wood, Sina Kafiabadi, Ayisha Al Busaidi, Emily Guilhem, Antanas Montvila, Jeremy Lynch, Matthew Townend, Siddharth Agarwal, Asif Mazumder, Gareth J Barker, et al. Accurate brain-age models for routine clinical MRI examinations. *Neuroimage*, 249:118871, 2022.
15. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
16. John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for? *Queue*, 6(2):40–53, 2008.
17. Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.
18. Anders Kallner. Resolution of Students t-tests, ANOVA and analysis of variance components from intermediary data. *Biochemia medica*, 27(2):253–258, 2017.
19. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
20. Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J Lund, Ingrid Agartz, Dag Alnæs, Deanna M Barch, Ramona Baur-Streubel, Alessandro Bertolino, et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623, 2019.
21. Patrick Foley, Micah J Sheller, Brandon Edwards, Sarthak Pati, Walter Riviera, Mansi Sharma, Prakash Narayana Moorthy, Shih-han Wang, Jason Martin, Parsa Mirhaji, et al. OpenFL: the open federated learning library. *Physics in Medicine & Biology*, 67(21):214001, 2022.
22. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
23. Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, Pedro PB de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.
24. Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017*, 2018.
25. T Uher, M Vaneckova, MP Sormani, J Krasensky, L Sobisek, J Blahova Dusankova, Z Seidl, E Havrdova, T Kalincik, RHB Benedict, et al. Identification of multiple sclerosis patients at highest risk of cognitive impairment using an integrated brain magnetic resonance imaging assessment approach. *European journal of neurology*, 24(2):292–301, 2017.

DRAFT