Structural brain damage and cognition in MS: an AI approach

Denissen, Stijn

Publication date:
2023

VRIJE
UNIVERSITEIT
BRUSSEL

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Medical Sciences

Structural brain damage and cognition in MS: an AI approach

Stijn Denissen

Academic year 2023-2024

Prof. Dr. ir. Guy Nagels
Prof. Dr. ir. Jeroen Van Schependom
Dr. Diana Sima

Faculty of Medicine and Pharmacy

AI-supported Modelling in Clinical Sciences (AIMS)

Printed by
Crazy Copy Center Productions
Vrije Universiteit Brussel
Pleinlaan 2
B-1050 Brussel
Tel: +32 2 629 33 44
E-mail: crazycopy@vub.be
www.crazycopy.be

ISBN: 9789464948066
NUR: 954
THEMA: MKJ, MKSG, UYQM

Supervisors

Prof. Dr. ir. Guy Nagels

- AI-supported Modelling in Clinical Sciences (AIMS) lab, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium
- Neurology Department, UZ Brussel, Brussels, Belgium
- St Edmund Hall, University of Oxford, Oxford, UK

Prof. Dr. ir. Jeroen Van Schependom

- AI-supported Modelling in Clinical Sciences (AIMS) lab, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium
- Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussels, Belgium

Dr. Diana Maria Sima

- icometrix, Leuven, Belgium
- AI-supported Modelling in Clinical Sciences (AIMS) lab, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium

External member PhD advisory board

Prof. Dr. Johan De Mey

- Radiology Department, UZ Brussel, Brussels, Belgium

Jury Members

Prof. Dr. Nicole Pouliart - Chair

- Department of Orthopaedics and Traumatology, Vrije Universiteit Brussel, UZ Brussel, Brussels, Belgium
- Department of basic (bio-)medical sciences, Vrije Universiteit Brussel, Brussels, Belgium

Prof. Dr. Letizia Leocani

- Università Vita-Salute, San Raffaele Hospital, Milan, Italy
- Experimental Neurophysiology Unit, Institute of Experimental Neurology (INSPE) - IRCCS-San Raffaele Hospital, Milan, Italy
- Department of Rehabilitation Sciences, Casa di Cura Igea, Milan, Italy

Prof. Dr. ir. Diego Vidaurre Henche

- Department of Clinical Medicine, Center of Functionally Integrative Neuroscience, Aarhus university, Aarhus, Denmark
- Department of Psychiatry, Oxford Centre for Human Brain Activity (OHBA), Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK

Prof. Dr. ir. Johan Loeckx

- Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium

Prof. Dr. Jan Versijpt

- Neurology Department, UZ Brussel, Brussels, Belgium
- Neuroprotection & Neuromodulation (NEUR), Vrije Universiteit Brussel, Brussels, Belgium

Contents

Acknowledgements

To my supervisors

I was lucky enough to have three promoters supervising my PhD trajectory. My two academic promoters from the VUB, Prof. Guy Nagels and Prof. Jeroen Van Schependom, together form a magnificent team in managing the AIMS lab. They do not scare away from providing us PhD students valuable memes, for example posing with a tiara in the Efteling, but most importantly complement each other in giving each of us the guidance we need. Guy and Jeroen, thank you for sharing my enthusiasm for all the various projects I have been involved in during my PhD, but simultaneously keep me focused on my research and help me overcome insecurities and manage perfectionism. I furthermore enjoyed sharing a passion for mindfulness with Guy, who was always there to comfort me by saying: "Relax, everything is out of control". Guy, I enjoyed teaching together, making bonfires in your garden during cold winter times and travelling together to start up the federated learning network in the last year of my PhD. In this light, I am especially grateful to you for personally bringing me to Prague with the brothers Laton and two powerful computers to kickstart the federated learning network. This road trip was a wonderful start of a three months research stay, where we shared a passion for submerging in the world of federated learning.

I admire my third supervisor, Dr. Diana Sima, for the ability to handle so many projects simultaneously at ico**metrix** with a great deal of passion, humour and positivity. Diana, even in your busy schedule, you always found time to patiently answer my questions, boost my self-confidence and guide me from an industrial and academic perspective. I loved going for a run together and discussing exciting new research avenues, which started with the Baekeland grant for which you have been an incredible support. I am confident that we will stay in touch in the future, for example with a business meeting during a refreshing run along the Vaart in Leuven.

To my jury members

I would like to express my sincere appreciation to the jury members of my thesis, Prof. Nicole Pouliart, Prof. Letizia Leocani, Prof. Diego Vidaurre, Prof. Johan Loeckx and Prof. Jan Versijpt. Thank you for investing time and energy in critically reviewing my PhD. Besides improving it, our discussions encouraged me to zoom out to the big picture in the final stage of my PhD.

To my blackbird

My deepest appreciation and respect go to my girlfriend, Merel Moens, who has the superpower to drag me out of my head and into the present moment. Sometimes people say that people who are most different are the ones that fit together best. Indeed, some of your interests could not be more different from mine. In this way, we drag each other out of our comfort zones, although we refuse to abandon the comfort zone that we have created together with our amazing cat Kamiel. I could not be more grateful for all your support during my PhD, enduring the highs and the lows, which I know has not been easy for you at times. Your caring, honest and positive character inspires me each day to be a better person. Thank you for being just you.

To my AIMS colleagues

In the early days of AIMS, when it was still called CIME, I loved playing futsal at campus Etterbeek with Lars and Johan and gaming on Lars's Play Station on the huge screen that Guy had bought to better read code together. Lars, you did not collect your Play Station to date, which really is just fine. You can rest assure that it will be put to good use in our new landscape office in Jette, although I hope you will pay a visit in the future to establish once and for all who the FIFA king is.

Johan, thank you for being my calm and happy office buddy in Etterbeek. You are the only one I have ever known to cycle from Brussels to Lier on a regular bike, and I challenge you to repeat it with a unicycle once. As the honest and caring person I know you are, I have no doubt that you will be an amazing dad for Rayén.

I could not have asked for a better ico**gnition** buddy than you, Delphine. I am proud of the study we conducted together, and for all the nice moments we shared during the process, such as recording the digit span in a bathroom since the acoustics were best there, and sharing a passion for running during working hours. You truly are an inexhaustible source of energy, and transfer this to every person you meet.

From teaching digital signal processing to Italian, Chiara, you are indeed a wonderful teacher. Moreover, you proved to be a great tour guide during our visit to Milan and Mantova, which I enjoyed to the fullest. I promise to never try out funny VPN constructions anymore without first notifying you, but will continue to convince you that the creature on the roof was a pigeon, not a duck.

If I were to be in an emergency computer situation, there are two persons I would call, and they happen to share their family name. Jorne and Jelle, there would be no federated learning network without you. Thank you for teaching me all the ins and outs of computer hardware and Linux. As there is so much left to learn, I am more than happy you will continue to be part of the AIMS intergalactic fellowship, which will continue its journey to Naples upcoming April.

Fahimeh, thank you for a wonderful PhD dinner last year. Saba, please keep sharing your amazing baking skills with us. Thomas, thanks for those much-needed moments of true bromance, and Gaia, thank you for your persistence in teaching me Italian. Frederik, thanks for sharing a passion for cycling with me during your time in our lab. Thank you AIMS, STIMULUS and NEUR colleagues for being awesome office buddies in Jette, I am looking forward to sharing many crazy moments in the future!

To my icolleagues

In 2019, I joined the ico**metrix** family, where I started preparing a Baekeland industrial grant application. I knew little about machine learning at the time, but could always count on my colleagues to patiently and enthusiastically explain me the concepts I wanted to know more about. The passion for AI research and its applications in medical imaging vibrates throughout ico**metrix**, which translates in the many milestones that have already been achieved. I am humbled to have had the opportunity to collaborate with, and most importantly learn from, so many bright, passionate and kind people throughout my PhD. To each and every one within ico**metrix**, thank you for the wonderful years, the offsites, the lunch runs, the beer tastings, Christmas breakfasts, the Italian dinners, "andare in bici" and all the other crazitivity I have had the pleasure of experiencing. Hopefully we can keep on collaborating in the future, and grab a beer from time to time in Leuven!

To all other collaborators

I would like to thank my dear colleagues at the National MS Center of Melsbroek for supporting my PhD trajectory from the start. Allowing me to present results on various occasions to different audiences, including patients with MS, gave me valuable new perspectives from a daily care point of view.

A big thanks to the radiology department of the UZ Brussel for allowing me to start my PhD in preparation of a personal grant, and for scanning my brain for a workshop for the Children's University of the VUB. Prof. Johan De Mey, thank you for your excellent guidance as external advisor of my PhD.

To the wonderful colleagues of the neurology department of the UZ Brussel: thank you for the nice vibes during my visits and for all the support you offered during my PhD.

To my dear colleagues in Prague: Minulý rok jsem prožil nádherné chvíle a vždycky budu vděčná za to, že jsem byl přivítá s otevřenou náručí. Děkuji za běhání, pozvání k vám domů a piva, která jsme společně vypili. Těším se na pokračování naší spolupráce i v budoucnu, a často vrátit do Prahy pro "řikat dobrý den"!

To Prof. Matthias Grothe and Robert Malinowski from Greifswald: Vielen Dank für das Vertrauen, das Sie mir von Beginn meiner Promotion an entgegengebracht haben, indem Sie mir erlaubt haben, mit Ihren Daten zu modellieren und unser Federated Learning Netzwerk zu unterstützen. Ich habe es genossen, Sie im letzten Jahr zweimal zu besuchen, und erinnere mich noch lebhaft daran, wie ich mit bloßen Füßen am Ostseestrand eine Fritz-Kola getrunken habe. Auf viele weitere solcher Treffen!

To my future colleagues in Naples: Vi ringrazio molto per la vostra fiducia. Non vedo l'ora di viaggiare da voi per iniziare la nostra collaborazione nel mese di aprile di quest'anno, e assaggiare se le pizze Napolitane sono davvero così buone come ho sentito dire!

To Prof. Iris-Katharina Penner: I enjoyed the road trip with Prof. Nagels to your centre in Düsseldorf, and would like to thank you for sharing data with our lab. I am looking forward to continue our collaboration in the future.

Alexander, thank you for your incredible patience during my first steps in the wondrous world of coding, while airplanes were arriving and taking off at Brussels Airport in the background.

Mireille, thank you for the wonderful lunches we had during my working days in Terhagen. I wish you the best of luck with your piano endeavours.

To all other researchers with whom I had the pleasure of collaborating in various studies: Thank you for the interesting discussions, your bright insights and humour!

To my family and friends

I can always count on unconditional love, support and confidence from my family in the Netherlands and Belgium. Oma, Pap, Mam, Niels, Jorinde, Renée, Gemma, Simone, Peter, Inge, Mitte and Jonathan, thank you for patiently listening to me, sharing happy and difficult moments and always encouraging me to chase my dreams, even if it would take me further from home. Thank you for printing my brain, drumming on my outreach songs and shaping ideas with me. You are the very best I could ever wish for.

To my dear friends in the Netherlands, Belgium and beyond. Thank you for having my back and sharing many moments of laughter, sports and music. Although distance complicates seeing some of you frequently, I every time enjoy picking up where we left off last time. To me, this indicates true friendship. Thanks for sharing it with me.

To the open source community

Thank you for sharing models, code, software and data, and publishing papers open access. Without your efforts, this thesis would not have been possible.

To all study participants

Thank you for your interest in our studies, your valuable feedback and taking the time to participate.

To Baudouin

Thank you for encouraging me to be a mindful PhD student, and to be more present and aware throughout life.

To whom I did not address

Thank you for being part of my PhD journey, and shaping the person that I am today.

Funding

This PhD was funded by the radiology department of the UZ Brussel, a Baekeland grant from VLAIO (HBC.2019.2579), an FWO travel grant (V412023N), an ECTRIMS travel grant and an IOF-POC grant from VUB.

Cover image

The cover image includes both a biological neural network (an axial slice of my brain) and an artificial neural network (AAN). The edges connecting the nodes of the AAN form triangles and quadrilaterals, of which 75% were filled randomly with colors used by Piet Mondriaan. The GitHub repository that creates this figure is named after him; MondrAIn (link to GitHub repository: QR code below).

I have named the creation "**A head full of perceptrons**". This is what I said I had to my promotor, Prof. Guy Nagels, when I was studying neural networks for a computer engineering course at the VUB. A perceptron is a unit of an AAN (cfr. chapter 4), and can be regarded as a mathematical representation of a biological neuron (chapter 1).

The image is a subtle reference to the elements on which my PhD is constructed. Additionally, the process of creating this figure illustrates my passion to combine science and art during my PhD by using songs and animations to do science communication. Lastly, by making the MondrAIn GitHub repository publicly available, I aim to say a word of thanks to the open science community that allowed me to perform my PhD research. Sharing data and code drives research and its transparency, which in my opinion is the way forward in any scientific discipline.

List of Figures

List of Tables

List of Abbreviations

aBDS	auditory Backwards Digit Span
ACTH	Adrenocorticotropic Hormone
AD	Alzheimer's Disease
ADL	Activities of Daily Living
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis Of Variance
APC	Antigen-Presenting Cell
BDI	Beck Depression Inventory
BICAMS	Brief International Cognitive Assessment for Multiple Sclerosis
BIDS	Brain Imaging Data Structure
BOLD	Blood-Oxygen Level Dependent
BPAD	Brain-Predicted Age Difference
BRB-N	Brief Repeatable Battery of Neuropsychological tests
BVMT-R	Brief Visuospatial Memory Test—Revised
CDSS	Clinical Decision Support System
CIS	Clinically Isolated Syndrome
CNN	Convolutional neural network
CNS	Central Nervous System
CSF	Cerebro-Spinal Fluid
CUDA	Compute Unified Device Architecture
CV	Cross-Validation
CVLT-II	California Verbal Learning Test—Second Edition
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DMT	Disease-Modifying Therapy
DP-SGD	Differentially Private Stochastic Gradient Descent
DSST	Digit-Symbol Substitution Test

DTI	Diffusion Tensor Imaging
DWI	Diffusion Weighted Imaging
EDSS	Expanded Disability Status Scale
EEG	Electro-Encephalography
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
FDA	Food and Drug Administration
FedAvg	Federated Averaging
FL	Federated Learning
FLAIR	Fluid-Attenuated Inversion Recovery
fMRI	Functional Magnetic Resonance Imaging
FSMC	Fatigue Scale for Motor and Cognitive Functions
GA	Glatiramer Acetate
GAN	Generative Adversarial Network
Gd	Gadolinium
GDPR	General Data Protection Regulation
GPU	Graphical Processing Unit
HC	Healthy Control
HLA	Human Leukocyte Antigen
ICC	Intraclass Correlation Coefficient
IPS	Information Processing Speed
IQR	Interquartile Range
IRT	Immune Reconstitution Therapy
LOGO	Leave One Group Out
MACFIMS	Minimal Assessment of Cognitive Function In MS
MAE	Mean Absolute Error
MAGNIMS	Magnetic Resonance Imaging In MS
MEG	Magneto-Encephalography
MHC	Major Histocompatibility Complex
ML	Machine Learning
MNI	Montreal Neurosciences Institute
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
MSE	Mean Squared Error
NEDA	No Evidence of Disease Activity
NIfTI	Neuroimaging Informatics Technology Initiative
NRMSE	Normalized Root Mean Squared Error
OCT	Optical Coherence Tomography

OUP	Oxford University Press
PCA	Principal Component Analysis
PPMS	Primary Progressive MS
PRMS	Progressive Relapsing MS
PRO	Patient-Reported Outcome
PROMS	Patient-Reported Outcome Measures
Q&A	Question and Answer
RCT	Randomized Controlled Trial
RF	Radio-Frequency
RF	Random Forest
RMSE	Root Mean Squared Error
RRMS	Relapsing-Remitting MS
S1PR	Sphingosine 1-Phosphate Receptor
SCP	Secure Copy Protocol
SD	Standard Deviation
SDMT	Symbol Digit Modalities Test
SGD	Stochastic Gradient Descent
sMRI	Structural Magnetic Resonance Imaging
SPART	Spatial Recall Test
SPMS	Secondary Progressive MS
SRT	Selective Reminding Test
SSH	Secure shell
SVM	Support Vector Machine
Th	T helper cell
TL	Transfer Learning
Treg	T-regulatory cell
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
vBDS	visual Backwards Digit Span
VPN	Virtual Private Network
WHO	World Health Organisation
XAI	Explainable AI

Preface

Cognitive problems are common in multiple sclerosis (MS). Yet even with advanced neuroimaging techniques, the origin of decline remains poorly understood. This thesis explores whether AI can shed new light on the structural underpinnings of cognitive impairment in MS. The story line below illustrates the coherence of its chapters.

About 3 million people worldwide are affected by multiple sclerosis (chapter 1), of which up to 70% have some form of cognitive decline (chapter 2). Although the link with structural brain damage has been studied intensively using magnetic resonance imaging (MRI, chapter 3), features describing the brain's structure fall short in explaining cognitive impairment. In my PhD, I investigated whether artificial intelligence (AI, chapters 4 and 5) could offer new insights. However, there is a catch. Machine learning usually requires large data sets, to which individual research labs do not have access. To still be able to pursue the investigation, I explored three solutions (hypotheses, chapter 6): 1) facilitating the collection of data with a smartphone-based cognitive screening battery (ico**gnition**, chapter 7), 2) reducing the need for data using the concept of brain age and transfer learning (chapters 8 and 9) and 3) exploring an alternative way to access data, namely with federated learning (chapter 9). After addressing the question whether AI will change MS care within the next 10 years (chapter 10), the thesis concludes with a discussion including potential future avenues (chapter 11).

Enjoy reading this manuscript that I have been shaping over the past 4 years. I like to think of it as my most exquisite piece of organised chaos. I invite the reader to keep in mind that this manuscript has come to being by a process that I believe to be the spark of any progress; the acceptance of change.

"Life is flux"
- Heraclitus

Summary

Multiple sclerosis (MS) affects about 3 million people worldwide, and is typically diagnosed in young adults. The disease affects the central nervous system, where it causes inflammation of nerve tissue and gradual degeneration of the nerve cells. The main target of MS is the insulation layer surrounding the nerve cells, reducing or even preventing signal transmission. The damage is visible on magnetic resonance (MR) images of the brain and spinal cord as regions of inflammation (lesions) and loss of brain tissue (atrophy). MS leads to a wide range of symptoms including cognitive impairment, which is present in up to 70% of people with MS.

Radiological findings however do not always manifest as clinical symptoms and the other way around, known as the "clinico-radiological paradox". The main goal of this thesis was to offer new insights in this paradox for cognitive problems using artificial intelligence (AI). However, databases with brain images and cognitive information are scarce, impeding AI research. This thesis proposes three solutions for this problem: (1) facilitating data collection with digital cognitive tests, (2) reducing the need for large databases by using models that are trained to perform a related task (transfer learning) and (3) increasing accessibility to clinical datasets for AI modelling using federated learning.

First, icognition was presented. This is a smartphone-based application with three cognitive tests, designed to screen for impairment in the two most commonly affected cognitive domains in MS: memory and information processing speed. The application was shown to be reliable and valid, although the results should be confirmed in a cognitively impaired MS sample. The application allows screening for cognitive problems at home, thereby picking up cognitive deterioration early on. Furthermore, digitalising cognitive tests facilitates the creation of large research databases in the future. Second, "brain age", interpretable as "how old the brain looks", was explored as an

in-between step to predict cognitive functioning. A model was trained to predict age, i.e. brain age, from brain MR images. The model overestimated age in people with MS, confirming that people with MS have older looking brains. Brain age moreover correlated with their information processing speed. Subsequently, a deep learning brain age model, trained on a large database of healthy people, was fine-tuned to predict cognition on a smaller MS database (transfer learning). Although the final model performed rather poorly at predicting cognitive performance from brain MRI, we proved in the same study that the model could be trained without sharing data between clinical centres. Instead of first collecting all data at one place, the model was sent to the data, where it was updated locally (federated learning). In this way, the model exhibited learning behaviour at each clinical centre, setting the stage for training better models without sharing sensitive clinical data such as brain MRI.

This thesis explored solutions for AI research in a context of low data availability. Reaching large and high-quality data sets could eventually enable AI to help patients with MS and their caregivers managing an unpredictable and burdensome disease.

Samenvatting

Multiple sclerose (MS) treft wereldwijd ongeveer 3 miljoen mensen en wordt meestal bij jongvolwassenen vastgesteld. De ziekte tast het centrale zenuwstelsel aan en veroorzaakt ontsteking van het zenuwweefsel en geleidelijke degeneratie van de zenuwcellen. Het belangrijkste doelwit van MS is de isolatielaag rond de zenuwcellen, waardoor de signaaloverdracht wordt verminderd of zelfs onmogelijk wordt gemaakt. De schade is zichtbaar op magnetische resonantie (MR) beelden van de hersenen en het ruggenmerg als gebieden van ontsteking (laesies) en verlies van hersenweefsel (atrofie). MS leidt tot een breed scala aan symptomen, waaronder cognitieve stoornissen, die bij tot 70% van de MS-patiënten voorkomen.

Radiologische bevindingen manifesteren zich echter niet altijd als klinische symptomen en andersom, bekend als de "klinisch-radiologische paradox". Het hoofddoel van dit proefschrift was om nieuwe inzichten te bieden in deze paradox voor cognitieve problemen met behulp van artificiële intelligentie (AI). Databases met hersenbeelden en cognitieve informatie zijn echter schaars, wat AI-onderzoek belemmert. Deze dissertatie stelt drie oplossingen voor dit probleem voor: (1) het vergemakkelijken van dataverzameling met digitale cognitieve testen, (2) het verminderen van de behoefte aan grote databases door modellen te gebruiken die getraind zijn om een gerelateerde taak uit te voeren (transfer learning) en (3) het vergroten van de toegankelijkheid van klinische datasets voor AI-modellering met behulp van federated learning.

Als eerste werd ico**gnition** gepresenteerd. Dit is een smartphone-applicatie met drie cognitieve tests, ontworpen om te screenen op beperkingen in de twee meest aangetaste cognitieve domeinen bij MS: geheugen en informatieverwerkingssnelheid. De applicatie bleek betrouwbaar en valide te zijn, hoewel de resultaten bevestigd moeten worden in een MS-groep met cognitieve beperkingen. De applicatie maakt het mogelijk om thuis te screenen op cognitieve problemen, waardoor cognitieve achteruitgang in een vroeg stadium kan worden

opgemerkt. Bovendien vergemakkelijkt het digitaliseren van cognitieve tests het creëren van grote onderzoeksdatabases in de toekomst. Ten tweede werd "hersenleeftijd", te interpreteren als "hoe oud de hersenen eruit zien", onderzocht als tussenstap om cognitief functioneren te voorspellen. Er werd een model getraind om leeftijd, d.w.z. hersenleeftijd, te voorspellen op basis van MR-beelden van de hersenen. Het model overschatte de leeftijd bij MS-patiënten, wat bevestigt dat MS-patiënten ouder uitziende hersenen hebben. De hersenleeftijd correleerde bovendien met hun informatieverwerkingssnelheid. Vervolgens werd een deep learning-model voor hersenleeftijd, getraind op een grote database van gezonde mensen, verfijnd om cognitie te voorspellen op een kleinere MS-database (transfer learning). Hoewel het uiteindelijke model vrij slecht presteerde in het voorspellen van cognitieve prestaties op basis van MRI van de hersenen, bewezen we in hetzelfde onderzoek dat het model getraind kon worden zonder gegevens uit te wisselen tussen klinische centra. In plaats van eerst alle gegevens op één plaats te verzamelen, werd het model naar de gegevens gestuurd, waar het lokaal werd getraind (federated learning). Op deze manier vertoonde het model leergedrag in elk klinisch centrum. In de toekomst kunnen hierdoor betere modellen worden getraind, zonder het delen van gevoelige klinische gegevens zoals MRI van de hersenen.

Deze dissertatie onderzocht oplossingen voor AI-onderzoek in een context van lage beschikbaarheid van data. Het bereiken van grote en hoogwaardige datasets zou AI uiteindelijk in staat kunnen stellen om patiënten met MS en hun zorgverleners te helpen bij het omgaan met een onvoorspelbare en belastende ziekte.

Part I

Background

Chapter 1

Multiple Sclerosis

Multiple sclerosis (MS) affects about 3 million people worldwide [1]. The disease is typically diagnosed in young adults and is characterised by a heterogeneous disease course. MS is an inflammatory and neurodegenerative disease that attacks the central nervous system (CNS), leading to the disintegration of myelin that serves to speed up signal conduction across nerve cells, as well as neuronal death. The inflammation is caused by an auto-immune reaction, in which the body's immune system starts attacking its own CNS. The inflammation itself can cause neurodegeneration, but neurodegeneration can also be present in the absence of inflammation, characterising the neurodegenerative component [2]. The eventual damage to the CNS is very heterogeneous in space and time, leading to difficulty in assigning an accurate prognosis to an individual patient with MS. Ultimately, the damage leads to impaired signal conduction, which gives rise to a plethora of clinical symptoms.

1.1 Symptoms

Due to damage done to the CNS, people with MS experience a wide range of symptoms. Some symptoms are more visible than others. For example, reduced strength, walking problems and impaired balance are typically observed [3]. Physical disability in MS is typically quantified by the Expanded Disability Status Scale (EDSS) [4]. There are however a multitude of other symptoms that cause significant inconvenience to people with MS, from bladder, bowel and sexual dysfunction to fatigue and cognitive problems (discussed in chapter 2) [3]. The collection of symptoms lead to activity restrictions such as carrying out many activities of daily living (ADL). Both the symptoms and activity restrictions can lead to reduced participation in society. Besides

objective problems, repercussions of MS involve subjective complaints such as a reduced quality of life [5] or subjective cognitive impairment [6].

1.2 The nervous system

The central nervous system (CNS) consists of the brain, the spinal cord and the optic nerves. The brain is the control centre of our body, whereas the spinal cord can be considered the highway of the nervous system; it allows fast signal conduction across the central nervous system. The other part of the nervous system is referred to as the peripheral nervous system, which conducts signals to (sensory signals, afferent neurons) and from (motor signals, efferent neurons) the CNS. The peripheral nervous system is furthermore divided in the somatic (voluntary) nervous system and the autonomous (involuntary) nervous system, the latter being involved in e.g. breathing, regulating blood pressure and digestion. The somatic nervous system [7] controls body movements via skeletal muscles.

1.2.1 The neuron and its function

One type of cells through which the nervous system communicates are neurons. It does so by transmitting signals, which are electrical by nature within each neuron, and neurochemical between neurons. The anatomy of a neuron (figure 1.1) is explained here by its function, namely conducting an electrical signal. First, each neuron contains dendrites that receive information from other neurons. These signals are subsequently collected in the soma (cell body), and if they exceed a certain threshold, a signal is generated for the next neurons: the action potential. This signal is in turn conducted by an axon to the synaptic cleft, which is the junction in between a neuron and the next. To be able to efficiently conduct an electrical signal, the axon of a nerve is covered by myelin that serves as an insulation layer. The myelin, created by oligodendrocytes, allows saltatoric signal conduction, which drastically speeds up transmission since the signal "jumps" across the axon.

1.2.2 MS affects the CNS

In MS, the immune system attacks the neurons in the CNS by targeting myelin and oligodendrocytes [8], which can lead to axonal degeneration. Axonal degeneration can however also happen independently of demyelination by other disease mechanisms in MS [9]. Altogether, the damage can lead to slowed

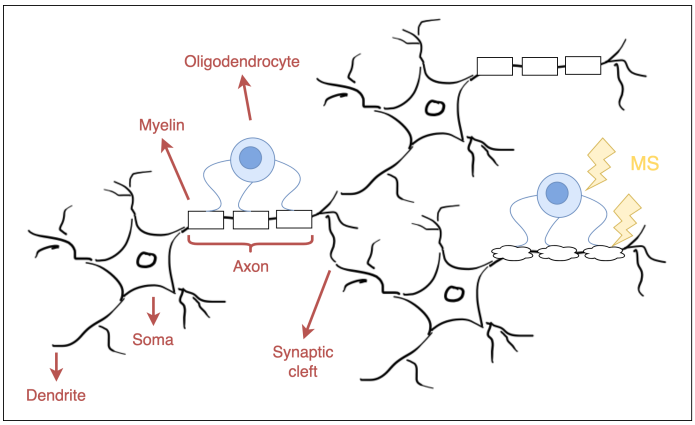

Figure 1.1: Three neurons in a biological neural network. The different components of the neuron, as well as the oligodendrocyte creating the myelin sheets, are indicated on the leftmost neuron. The right lower neuron illustrates the damage done by MS, attacking the myelin sheet and oligodendrocytes [8].

signal conduction or prevent the signal from passing at all. The next section elaborates on the inflammatory component of MS.

1.3 A flaw in the immune system

Humans possess an impressive mechanism to defend the body against external and internal threats; the immune system. It consists of a first-line defence mechanism (the innate immune system) and the adaptive immune system, that as the name suggest can adapt to specific threats [10]. The innate immune system firstly consists of anatomical barriers (e.g. skin) that aim to keep unwanted particles from invading the body. Whenever particles still find their way inside the body, the innate immune system is equipped with other defences. It will try to make the living conditions for the particles undesirable, increase the defence capability of uninfected cells, attack the intruders directly, call for aid and facilitate transport of this aid. The innate immune system can respond quickly, but is less specialised to specific threats compared to the adaptive immune system. Although it reacts slower, the adaptive immune system is able to memorise previous attacks, and is therefore more efficient for handling specific threats [10].

The cellular component of the adaptive immune system consists of two types of cells; B cells and T cells. Both are triggered by the presence of a

foreign antigen, for example microbes, viruses and toxins, but also cancer cells [10]. B cells can directly recognise an antigen, and respond by either proliferating into plasma cells that contain antibodies that bind to the antigen to flag the cell for destruction [11], or into memory B cells to respond quicker to a future encounter with the antigen. T cells need so-called antigen-presenting cells (APC) to recognise an antigen, a process that is nicely illustrated in Kasper and Shoemaker 2010 [12]. In case the T cell is a $CD8^+$ T cell, it will proliferate in a cytotoxic T cell, whose main function is cell destruction [10]. In case the T cell is a $CD4^+$ T cell, it will proliferate primarily in one of three different T helper cells (Th), whose function is mediating the immune response; Th1, Th2 and Th17 [10]. Th1 and Th17 promote inflammation, whereas Th2 reduces inflammation [3, 11]. Another type of $CD4^+$ T cell is a T-regulatory cell (Treg), which has a role in regulating the immune response. When the immune response resolves, only a fraction of the T cells remain as memory cells for a quicker response in the future [10].

In some cases, the immune system is not properly regulated, which can lead to inefficiency (immunodeficiency), overreaction (hypersensitivity), or even mistaking healthy parts of the own body for threats (autoimmunity) [10]. In MS, this autoimmune reaction targets the central nervous system (CNS). The disruption of the immune system in MS is complex, and according to Loma et al. 2011, involves both the innate and adaptive immune system [11]. The problem appears to be mainly on the level of T cells [13]. Although it is unknown what exactly triggers the disease, genetic predisposition and environmental exposure appear to play a role [3, 11].

1.4 Who is at risk of developing MS?

1.4.1 Environmental risk factors

In a review by Young et al. 2011 [14], several environmental risk factors are described. People living at a higher latitude seem to be at an increased risk of developing MS. This is also the case for people infected by the Epstein Barr Virus, leading to glandular fever. Other examples of risk factors are those related to sunlight exposure and vitamin D, where decrease is associated with an increased risk of developing MS [14, 15, 16].

1.4.2 Genetic risk factors

Having family members with MS raises the risk of acquiring MS [15]. More specifically, certain human leukocyte antigen (HLA) genes are associated with an increased risk [14]. HLA genes are important in the immune response as the proteins they encode help in deciding if a cell belongs to the body (self) or not (non-self) [17].

1.4.3 Other risk factors

Finally, a wide range of lifestyle-related factors such as obesity and smoking have been related to the development of MS [14, 16]. A relationship with migraine has been reported as well [15, 18].

1.5 Diagnosis

The diagnosis of MS relies on the McDonald criteria, of which the most recent version dates back to 2017 [19]. The diagnosis of MS is mainly assigned based on clinical examination, magnetic resonance imaging (MRI, chapter 3) and laboratory tests (e.g. analysis of the cerebrospinal fluid) [20]. The damage done to the CNS by MS can for example be observed clinically by the symptoms described above, and as inflammatory plaques and atrophy on MR images of the brain and the spinal cord. For the diagnosis to be made, there essentially has to be objective evidence that the CNS is attacked in a diffuse way (dissemination in space), which is not restricted to a single occurrence in time (dissemination in time). Dissemination in time can be different, with multiple episodes of disease exacerbation and (partial) recovery (relapsing-remitting MS, RRMS) or a disease process that is more gradual over time. The latter is referred to as primary progressive MS (PPMS) if this gradual process is present from the start, or secondary progressive MS (SPMS) if this occurs after RRMS. When the phenotypes were established in 1996, a fourth, more rare type was distinguished, involving a gradual disease activity with occasional relapses [21]. This type was referred to as progressive relapsing MS (PRMS), but has been rejected after redefining the subtypes in 2013 [22]. In the same revisions, clinically isolated syndrome (CIS) was acknowledged as an MS phenotype, which is a condition that clinically resembles MS, but its recurrence over time needs to be established [22].

1.6 Treatment

Although there is no cure for MS, different treatment options are available nowadays for managing the disease.

1.6.1 Disease-modifying therapy

Disease-modifying therapy (DMT), as the name suggests, aims to target the disease process underlying symptoms. They typically act upon a specific component of the immune system and alleviate symptoms by suppressing specific elements of the disease process. The majority of DMTs are approved only for people with RRMS [23], of which the most commonly prescribed ones are summarised in a brief overview below. The overview is based on Hauser and De Cree 2020 and Liu et al. 2021 [23, 24].

First-line treatment

Those are predominantly immunomodulators (medication that acts upon the immune system), and are focused on safety at the expense of efficacy [24].

- Dimethyl fumarate (Tecfidera®) exerts an anti-inflammatory effect by stimulating a Th2 response rather than a Th1 response [25]. Moreover, cytoprotective effects have been reported [23, 25], which could provide protection to cells of the CNS [25].

- Teriflunomide (Aubagio®) suppresses the proliferation of both B and T lymphocytes, thus modulating the immune system [26].

- Interferon beta (e.g. Avonex®, Plegridy®, Rebif®) has multiple effects, including a reduction of major histocompatibility complex (MHC) presentation on APCs (reducing antigen stimulation), stimulation and suppression of pro- and anti-inflammatory cytokines respectively (shift from a Th1/Th17 to a Th2 response), reducing proliferation of T-cells and preventing inflammatory cells from entering the CNS [27].

- Glatiramer acetate (GA, e.g. Copaxone®) is an amino acid polymer that resembles myelin basic protein (MBP), which is part of the myelin surrounding axons in the CNS. Antibodies in MS target MBP, who remain spared if the antibodies attack GA instead [28].

Second-line treatment

This group mainly consists of immuno-suppressiva and antibodies acting upon immune cells [24].

- The first group consists of antibodies that bind to the CD20 antigen on the cell surface of B cells, selectively depleting them. Ocrelizumab (Ocrevus®) is a frequently prescribed medication in this group.
- Other DMTs work by reducing the number of lymphocytes that reach the CNS. For example, Natalizumab (Tysabri®) inhibits $\alpha 4\beta 1$ integrin that is present on the cell surface of lymphocytes, thereby preventing migration across the endothelium. Alternatively, Fingolimod (Gilenya®) acts upon sphingosine 1-phosphate receptors (S1PR) on lymphocytes [29]. S1PR is subsequently internalised by the cell, preventing the lymphocyte from leaving a lymph node.
- Cladribine (Mavenclad®) and alemtuzumab (Lemtrada®) are considered a second-line drug [24], although both were also previously used as first-line treatment in some countries [30, 31]. They are immune reconstitution therapy (IRT), which aim is to act upon the immune system in a long-lasting way by using short treatment periods [30]. Cladribine and alemtuzumab deplete both B-cells and T-cells [30].

For progressive forms of MS, treatment options are far more limited, with ocrelizumab being the only DMT for PPMS [23]. For SPMS, siponimod (Mayzent®) is available, which is a $S1PR_1$ and $S1PR_5$ modulator [32]. However, there appears to be an active investigation ongoing to enrich the palette of treatment options [33].

1.6.2 Symptomatic treatment

Contrary to DMTs, the core focus of symptomatic treatment is not to address the disease mechanisms that cause the symptoms. Instead, they are prescribed specifically to alleviate symptoms, thereby reducing discomfort to the person suffering from them. The book chapter by Toosy et al. 2014 [34] nicely reflects the wide range of symptoms that people with MS experience (cfr. supra for a non-exhaustive overview of symptoms), and how to address them. Among treatment strategies are drug therapy (such as baclofen to reduce spasticity) and physical or cognitive rehabilitation to e.g. alleviate pain and spasticity or treat motor dysfunction [34, 35].

1.6.3 Relapse treatment

Lastly, episodes of disease exacerbation (relapses) are typically treated with corticosteroids or sometimes adrenocorticotropic hormone (ACTH) [36, 37]. Plasma exchange is sometimes used in patients that are resistant to steroids [38].

1.7 Prognosis

The disease course of MS is highly heterogeneous [39]. A plethora of efforts have emerged the past few decades in finding biomarkers that are suggestive of a distinct disease course (cfr. chapter 4). Models that combine biomarkers have also been considered to enhance predictive accuracy, the majority using fairly simple statistical models [40, 41]. More recent research investigates the potential of artificial intelligence (AI, chapter 4) to both uncover new biomarkers [42] and combine them to yield predictive models [43, 44].

The reality in clinical practice is that there is no model yet, simple or complicated, that can inform clinicians on the future course of the patients they are treating, let alone which treatment they can best prescribe to maximally reduce CNS damage. This too is an interesting avenue for AI research. Chapter 5 reviews the literature that uses AI to predict future cognitive deterioration in MS.

References

[1] Walton, C., King, R., Rechtman, L., Kaye, W., Leray, E., Marrie, R.A., Robertson, N., La Rocca, N., Uitdehaag, B., van Der Mei, I. et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS. *Multiple Sclerosis Journal*, 26(14):1816–1821, 2020.

[2] Sandi, D., Fricska-Nagy, Z., Bencsik, K. and Vécsei, L. Neurodegeneration in Multiple Sclerosis: Symptoms of Silent Progression, Biomarkers and Neuroprotective Therapy—Kynurenines Are Important Players. *Molecules*, 26(11):3423, 2021.

[3] Ghasemi, N., Razavi, S. and Nikzad, E. Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell Journal (Yakhteh)*, 19(1):1, 2017.

[4] Kurtzke, J.F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1452, November 1983.

[5] Papuć, E. and Stelmasiak, Z. Factors predicting quality of life in a group of Polish subjects with multiple sclerosis: accounting for functional state, socio-demographic and clinical factors. *Clinical neurology and neurosurgery*, 114(4):341–346, 2012.

[6] Jelinek, P., Simpson Jr, S., Brown, C., Jelinek, G., Marck, C., De Livera, A., O'Kearney, E., Taylor, K., Neate, S. and Weiland, T. Self-reported cognitive function in a large international cohort of people with multiple sclerosis: associations with lifestyle and other factors. *European Journal of Neurology*, 26(1):142–154, 2019.

[7] Akinrodoye, M.A. and Lui, F. Neuroanatomy, somatic nervous system. 2020.

[8] Zhao, X. and Jacob, C. Mechanisms of Demyelination and Remyelination Strategies for Multiple Sclerosis. *International Journal of Molecular Sciences*, 24(7):6373, 2023.

[9] Haines, J.D., Inglese, M. and Casaccia, P. Axonal damage in multiple sclerosis. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 78(2):231–243, 2011.

[10] Marshall, J.S., Warrington, R., Watson, W. and Kim, H.L. An introduction to immunology and immunopathology. *Allergy, Asthma & Clinical Immunology*, 14(2):1–10, 2018.

[11] Loma, I. and Heyman, R. Multiple sclerosis: pathogenesis and treatment. *Current neuropharmacology*, 9(3):409–416, 2011.

[12] Kasper, L.H. and Shoemaker, J. Multiple sclerosis immunology: the healthy immune system vs the MS immune system. *Neurology*, 74(1 Supplement 1):S2–S8, 2010.

[13] Huang, W.J., Chen, W.W. and Zhang, X. Multiple sclerosis: Pathology, diagnosis and treatments. *Experimental and therapeutic medicine*, 13(6):3163–3166, 2017.

[14] Young, C. Factors predisposing to the development of multiple sclerosis. *QJM: An International Journal of Medicine*, 104(5):383–386, 2011.

[15] Taan, M., Al Ahmad, F., Ercksousi, M.K. and Hamza, G. Risk factors associated with multiple sclerosis: a case-control study in Damascus, Syria. *Multiple Sclerosis International*, 2021:1–5, 2021.

[16] Alfredsson, L. and Olsson, T. Lifestyle and environmental factors in multiple sclerosis. *Cold Spring Harbor perspectives in medicine*, 9(4), 2019.

[17] Nordquist, H. and Jamil, R.T. Biochemistry, HLA Antigens. 2019.

[18] Mrabet, S., Wafa, M. and Giovannoni, G. Multiple sclerosis and migraine: Links, management and implications. *Multiple Sclerosis and Related Disorders*, page 104152, 2022.

[19] Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S. et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2):162–173, 2018.

[20] Ford, H. Clinical presentation and diagnosis of multiple sclerosis. *Clinical Medicine*, 20(4):380, 2020.

[21] Lublin, F.D., Reingold, S.C. et al. Defining the clinical course of multiple sclerosis: results of an international survey. *Neurology*, 46(4):907–911, 1996.

[22] Lublin, F.D., Reingold, S.C., Cohen, J.A., Cutter, G.R., Sørensen, P.S., Thompson, A.J., Wolinsky, J.S., Balcer, L.J., Banwell, B., Barkhof, F. et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, 83(3):278–286, 2014.

[23] Hauser, S.L. and Cree, B.A. Treatment of multiple sclerosis: a review. *The American journal of medicine*, 133(12):1380–1390, 2020.

[24] Liu, Z., Liao, Q., Wen, H. and Zhang, Y. Disease modifying therapies in relapsing-remitting multiple sclerosis: a systematic review and network meta-analysis. *Autoimmunity Reviews*, 20(6):102826, 2021.

[25] Bomprezzi, R. Dimethyl fumarate in the treatment of relapsing–remitting multiple sclerosis: an overview. *Therapeutic advances in neurological disorders*, 8(1):20–30, 2015.

[26] Oh, J. and O'Connor, P.W. Teriflunomide. *Neurology: Clinical Practice*, 3(3):254–260, 2013.

[27] Filipi, M. and Jack, S. Interferons in the treatment of multiple sclerosis: a clinical efficacy, safety, and tolerability update. *International journal of MS care*, 22(4):165–172, 2020.

[28] Babaesfahani, A. and Bajaj, T. Glatiramer. 2019.

[29] McGinley, M.P. and Cohen, J.A. Sphingosine 1-phosphate receptor modulators in multiple sclerosis and other conditions. *The Lancet*, 398(10306):1184–1194, 2021.

[30] Giovannoni, G. and Mathews, J. Cladribine Tablets for Relapsing–Remitting Multiple Sclerosis: A Clinician's Review. *Neurology and therapy*, 11(2):571–595, 2022.

[31] Berger, T., Elovaara, I., Fredrikson, S., McGuigan, C., Moiola, L., Myhr, K.M., Oreja-Guevara, C., Stoliarov, I. and Zettl, U.K. Alemtuzumab use in clinical practice: recommendations from European multiple sclerosis experts. *CNS drugs*, 31:33–50, 2017.

[32] Scott, L.J. Siponimod: a review in secondary progressive multiple sclerosis. *CNS drugs*, 34:1191–1200, 2020.

[33] Ciotti, J.R. and Cross, A.H. Disease-modifying treatment in progressive multiple sclerosis. *Current treatment options in neurology*, 20:1–26, 2018.

[34] Toosy, A., Ciccarelli, O. and Thompson, A. Symptomatic treatment and management of multiple sclerosis. *Handbook of clinical neurology*, 122:513–562, 2014.

[35] Kubsik-Gidlewska, A.M., Klimkiewicz, P., Klimkiewicz, R., Janczewska, K. and Woldańska-Okońska, M.Z. Rehabilitation in multiple sclerosis. *Advances in Clinical and Experimental Medicine*, 26(4), 2017.

[36] Berkovich, R. Treatment of acute relapses in multiple sclerosis. *Translational Neuroimmunology in Multiple Sclerosis*, pages 307–326, 2016.

[37] Repovic, P. Management of multiple sclerosis relapses. *CONTINUUM: Lifelong Learning in Neurology*, 25(3):655–669, 2019.

[38] Bunganic, R., Blahutova, S., Revendova, K., Zapletalova, O., Hradilek, P., Hrdlickova, R., Ganesh, A., Cermakova, Z., Bar, M. and Volny, O. Therapeutic plasma exchange in multiple sclerosis patients with an aggressive relapse: an observational analysis in a high-volume center. *Scientific Reports*, 12(1):18374, 2022.

[39] Lucchinetti, C., Brück, W., Parisi, J., Scheithauer, B., Rodriguez, M. and Lassmann, H. Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 47(6):707–717, 2000.

[40] Dekker, I., Eijlers, A., Popescu, V., Balk, L., Vrenken, H., Wattjes, M., Uitdehaag, B., Killestein, J., Geurts, J., Barkhof, F. et al. Predicting clinical progression in multiple sclerosis after 6 and 12 years. *European journal of neurology*, 26(6):893–902, 2019.

[41] Eijlers, A.J., van Geest, Q., Dekker, I., Steenwijk, M.D., Meijer, K.A., Hulst, H.E., Barkhof, F., Uitdehaag, B.M., Schoonheim, M.M. and Geurts, J.J. Predicting cognitive decline in multiple sclerosis: a 5-year follow-up study. *Brain*, 141(9):2605–2618, 2018.

[42] Kopf, A. and Claassen, M. Latent representation learning in biology and translational medicine. *Patterns*, 2(3), 2021.

[43] Seccia, R., Romano, S., Salvetti, M., Crisanti, A., Palagi, L. and Grassi, F. Machine Learning Use for Prognostic Purposes in Multiple Sclerosis. *Life 2021, Vol. 11, Page 122*, 11(2):122, feb 2021.

[44] Denissen, S., Chén, O.Y., De Mey, J., De Vos, M., Van Schependom, J., Sima, D.M. and Nagels, G. Towards multimodal machine learning prediction of individual cognitive evolution in multiple sclerosis. *Journal of Personalized Medicine*, 11(12):1349, 2021.

Chapter 2

Cognitive impairment

As outlined in chapter 1, people with MS experience a wide range of symptoms. The most visible among those symptoms are physical impairments such as gait difficulties and balance disturbances [1]. A symptom that is often more subtle but equally disabling is cognitive impairment, which is present in about half of people with MS [2]. Cognitive problems have negative repercussions on daily life activities [3] and quality of life [4], while on a societal level, it leads to an increased cost [5], which might be caused by a reported loss of employment and productivity [6].

2.1 Cognition

The complexity of what entails "cognition" is reflected by an article by Bayne et al. 2019 [7]. Eleven different experts were asked to share their view on cognition, leading to very distinct answers. The abstractness of the concept is reflected by the name; "together" (Latin: com) and "to know" (Latin: gnoscere) [7].

Cognition is observable as the behavioural output of intermediary processing between a sensation and a subsequent action [8]. According to Mesulam et al. 1998, the path between sensory input and action has evolved towards an optimum between short response time (short neuronal path length) and complex processing (longer path length and parallel processing), allowing the same stimulus to be processed differently, thus leading to different outcomes [8]. For example, being touched on the shoulder in the dark when not anticipated could lead to a startle reaction, while the same touch on the shoulder could lead to relaxation in different circumstances. The intermediate

processing steps are typically categorised as cognitive domains, and depending on the conceptualisation, could be hierarchical by nature. In the latter view, cognitive domains are characterised by increasingly complex processing, ranging from basic sensory processes to executive function. This however is no one-way traffic; higher level cognitive domains can exert feedback to lower cognitive domains [9]. A comprehensive overview of all cognitive domains can be found in Harvey et al. 2019 [9].

2.2 Cognitive domains affected by MS

Literature indicates that in MS, the two most commonly affected domains are memory and information processing speed [2, 10], and most cognitive tests to screen for them are also designed to measure these domains (cfr. infra). This however is a simplification of the complex palette of cognitive domains that appear to be affected in MS, such as several subdomains of memory and other domains like executive function and attention [2].

2.3 Biomarkers of cognitive impairment

To study cognitive impairment, it is useful to work with the concept of "biomarkers". Although there are multiple definitions for a biomarker, we will work here with the one proposed by the World Health Organisation (WHO): "almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction" [11]. In the context of cognitive impairment, the potential hazard constitutes cognitive deterioration. Searching for these characteristics has been made possible by technological advances such as electro- and magneto-encephalography (EEG and MEG respectively), magnetic resonance imaging (MRI, chapter 3) and analysis of cerebro-spinal fluid. Each of those can provide unique biological information. EEG and MEG for example measure brain function, whereas MRI can be used to assess both brain structure (sMRI) and brain function (fMRI). Each modality has both strengths and weaknesses; EEG for example has a high temporal resolution, meaning that fluctuations of brain signals can be measured much more detailed over time, but at the cost of spatial resolution, meaning that the source of the signal is hard to pinpoint. The inverse is true for fMRI.

As a result of applying the aforementioned technological advances to people with MS, a wide range of biomarkers for different cognitive domains have been identified. Indicators of neurodegeneration such as (regional) brain volume (MRI) [12] and Tau protein (cerebro-spinal fluid (CSF)) [13] for example correlate with information processing speed, as do MRI markers of neuroinflammation [14]. Also biomarkers related to brain function have been reported [15]. It is important to underline the importance of functional assessments such as MEG in the study of cognition in MS [16, 17, 18].

Up until now, a so-called "knowledge-based" representation was made from the raw data, namely by extracting summarising variables that are deemed important to assess cognitive impairment. With this transformation, we however risk to lose a lot of information. Artificial intelligence (AI, chapter 4) might overcome this issue by learning a "data-driven" representation of the raw data with regard to cognitive impairment using a technique called "deep learning". In this way, it might identify "hidden", "data-driven" biomarkers of cognitive impairment in the raw data, in turn revealing new insights in the biological underpinnings of cognitive impairment.

2.4 Assessment of cognitive impairment

Cognitive domains are usually assessed with so-called paper-pencil tests, and are often combined in cognitive batteries aiming to provide a bigger picture of a patient's cognitive status. Popular batteries include the Brief Repeatable Battery of Neuropsychological tests (BRB-N) by Stephen Rao [19], the minimal assessment of cognitive function in MS (MACFIMS) [20] and the Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) [21]. The BICAMS is the most recent of the three and was designed to screen for problems with IPS and memory. Particularly screening for IPS appears to be of interest, as it might be the first cognitive domain deteriorating in MS [22]. The most popular test for this domain is the Symbol Digit Modalities Test (SDMT) [23], of which a mock version is shown in figure 2.1.

In an ever-digitalising world, many research groups proposed computerised versions of these tests. They are predominantly digital versions of the SDMT and designed for tablets [25] or smartphones [26]. The advantages of using digital tests are mainly the following:

1. Tests can be completed individually at home, without an external rater

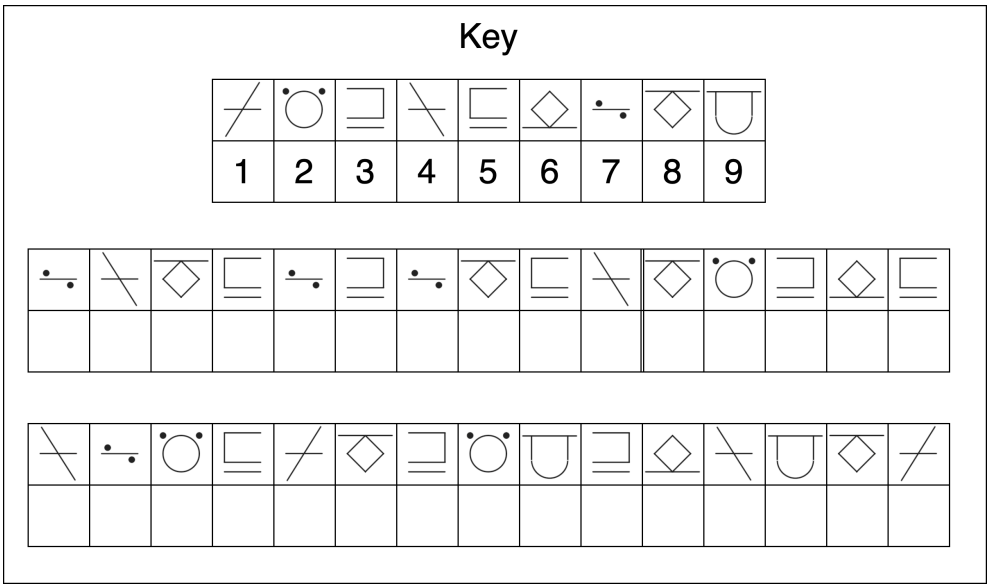

Figure 2.1: A shortened mock version of the symbol digit modalities test (SDMT) [23], using symbols from the ico**gnition** application (chapter 7) [24]. The subject is asked to convert each symbol to its corresponding digit using the key on top, one symbol at a time and saying the answer out loud. The goal is to convert as many symbols as possible in 90 seconds. Here, only 10 trial symbols (up until the double vertical line) and 20 test symbols are shown.

2. Solution keys can be randomised to prevent memorising a test

3. Testing can be done more frequently, increasing temporal resolution

4. Data is stored digitally, facilitating data analysis and modelling

In chapter 7, ico**gnition** is introduced, which is a recently developed smartphone-based cognitive screening battery. The chapter contains the validation study of the application, assessing IPS and memory [24]

2.5 How cognitive problems are treated

According to De Luca et al. 2020 [3], there is insufficient evidence for any pharmacological agent to treat cognitive impairment. Disease modifying therapy (DMT) appears not to target it directly, while symptomatic treatment such as dalfampridine (Fampyra®) might improve information processing speed, al-

though evidence is conflicting [3].

Cognitive rehabilitation however gains a lot of attention, mostly on the domain of learning and memory. The used interventions are heterogeneous, but primarily target a specific domain or use a multimodal approach. For both approaches, positive effects have been reported in a wide range of cognitive domains such as information processing speed, (working) memory and executive functions [3].

Beyond cognitive rehabilitation, also physical rehabilitation appears to have beneficial effects on cognitive problems in MS [3]. An on-going trial is currently investigating whether cognitive or physical rehabilitation as stand-alone treatment or a combination of both has superior effects on multiple cognitive domains in MS [27].

References

[1] Ghasemi, N., Razavi, S. and Nikzad, E. Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell Journal (Yakhteh)*, 19(1):1, 2017.

[2] Macías Islas, M.A. and Ciampi, E. Assessment and Impact of Cognitive Impairment in Multiple Sclerosis: An Overview. *Biomedicines*, 7(1):22, March 2019. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[3] DeLuca, J., Chiaravalloti, N.D. and Sandroff, B.M. Treatment and management of cognitive dysfunction in patients with multiple sclerosis. *Nature Reviews Neurology*, 16(6):319–332, 2020.

[4] Nabizadeh, F., Balabandian, M., Rostami, M.R., Owji, M., Sahraian, M.A., Bidadian, M., Ghadiri, F., Rezaeimanesh, N. and Moghadasi, A.N. Association of cognitive impairment and quality of life in patients with multiple sclerosis: A cross-sectional study. *Current Journal of Neurology*, 21(3):144, 2022.

[5] Maltby, V.E., Lea, R.A., Reeves, P., Saugbjerg, B. and Lechner-Scott, J. Reduced cognitive function contributes to economic burden of multiple sclerosis. *Mult Scler Relat Disord*, 60:103707, April 2022.

[6] Rodriguez Llorian, E., Zhang, W., Khakban, A., Michaux, K., Patten, S., Traboulsee, A., Oh, J., Kolind, S., Prat, A., Tam, R. et al. Employment status, productivity loss, and associated factors among people with multiple sclerosis. *Multiple Sclerosis Journal*, page 13524585231164295, 2023.

[7] Bayne, T., Brainard, D., Byrne, R.W., Chittka, L., Clayton, N., Heyes, C., Mather, J., Ölveczky, B., Shadlen, M., Suddendorf, T. et al. What is cognition? *Current Biology*, 29(13):R608–R615, 2019.

[8] Mesulam, M.M. From sensation to cognition. *Brain: a journal of neurology*, 121(6):1013–1052, 1998.

[9] Harvey, P.D. Domains of cognition and their assessment. *Dialogues in clinical neuroscience*, 21(3):227–237, 2019.

[10] Leavitt, V.M., Tosto, G. and Riley, C.S. Cognitive phenotypes in multiple sclerosis. *Journal of neurology*, 265:562–566, 2018.

[11] Strimbu, K. and Tavel, J.A. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

[12] Ziccardi, S., Pizzini, F.B., Guandalini, M., Tamanti, A., Cristofori, C. and Calabrese, M. Making Visible the Invisible: Automatically Measured Global and Regional Brain Volume Is Associated with Cognitive Impairment and Fatigue in Multiple Sclerosis. *Bioengineering*, 10(1):41, 2022.

[13] Virgilio, E., Vecchio, D., Crespi, I., Puricelli, C., Barbero, P., Galli, G., Cantello, R., Dianzani, U. and Comi, C. Cerebrospinal fluid biomarkers and cognitive functions at multiple sclerosis diagnosis. *Journal of Neurology*, 269(6):3249–3257, 2022.

[14] Pham, L., Harris, T., Varosanec, M., Morgan, V., Kosa, P. and Bielekova, B. Smartphone-based symbol-digit modalities test reliably captures brain damage in multiple sclerosis. *npj Digit. Med.*, 4(1):1–13, February 2021. Number: 1 Publisher: Nature Publishing Group.

[15] Khan, H., Sami, M. and Litvak, V. The utility of Magnetoencephalography in multiple sclerosis–A systematic review. *NeuroImage: Clinical*, 32:102814, 2021.

[16] Akbarian, F., Rossi, C., Costers, L., D'hooghe, M.B., D'haeseleer, M., Nagels, G. and Van Schependom, J. The spectral slope as a marker of excitation/inhibition ratio and cognitive functioning in multiple sclerosis. *Human Brain Mapping*, 2023.

[17] Rossi, C., Vidaurre, D., Costers, L., Akbarian, F., Woolrich, M., Nagels, G. and Van Schependom, J. A data-driven network decomposition of the temporal, spatial, and spectral dynamics underpinning visual-verbal working memory processes. *Communications Biology*, 6(1):1079, 2023.

[18] Costers, L., Van Schependom, J., Laton, J., Baijot, J., Sjøgård, M., Wens, V., De Tiège, X., Goldman, S., D'Haeseleer, M., D'hooghe, M.B. et al. The role of hippocampal theta oscillations in working memory impairment in multiple sclerosis. *Human brain mapping*, 42(5):1376–1390, 2021.

[19] Rao, S.M., Leo, G.J., Bernardin, L. and Unverzagt, F. Cognitive dysfunction in multiple sclerosis.: I. Frequency, patterns, and prediction. *Neurology*, 41(5):685–691, 1991.

[20] Benedict, R.H., Cookfair, D., Gavett, R., Gunther, M., Munschauer, F., Garg, N. and Weinstock-Guttman, B. Validity of the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society*, 12(4):549–558, 2006.

[21] Langdon, D., Amato, M., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hämäläinen, P., Hartung, H.P., Krupp, L., Penner, I. et al. Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *Mult Scler*, 18(6):891–898, June 2012.

[22] Van Schependom, J., D'hooghe, M.B., Cleynhens, K., D'hooge, M., Haelewyck, M.C., De Keyser, J. and Nagels, G. Reduced information processing speed as primum movens for cognitive decline in MS. *Mult Scler*, 21(1):83–91, January 2015.

[23] Benedict, R.H., DeLuca, J., Phillips, G., LaRocca, N., Hudson, L.D. and Rudick, R. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. *Mult Scler*, 23(5):721–733, April 2017.

[24] Denissen, S., Van Laethem, D., Baijot, J., Costers, L., Descamps, A., Van Remoortel, A., Van Merhaegen-Wieleman, A., D'hooghe, M.B., D'Haeseleer, M., Smeets, D. et al. icognition: a smartphone-based cognitive screening battery. *medRxiv*, pages 2023–07, 2023.

[25] Beier, M., Alschuler, K., Amtmann, D., Hughes, A., Madathil, R. and Ehde, D. iCAMS: Assessing the Reliability of a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) Tablet Application. *International Journal of MS Care*, 22(2):67–74, March 2020.

[26] Foong, Y.C., Bridge, F., Merlo, D., Gresle, M., Zhu, C., Buzzard, K., Butzkueven, H. and van der Walt, A. Smartphone monitoring of cognition in people with multiple sclerosis: A systematic review. *Multiple Sclerosis and Related Disorders*, page 104674, 2023.

[27] Van Laethem, D., Van de Steen, F., Kos, D., Naeyaert, M., Van Schuerbeek, P., D'Haeseleer, M., D'Hooghe, M.B., Van Schependom, J. and Nagels, G. Cognitive-motor telerehabilitation in multiple sclerosis (CoMoTeMS): study protocol for a randomised controlled trial. *Trials*, 23(1):1–10, 2022.

Chapter 3

Magnetic Resonance Imaging

About 50 years ago, Peter Mansfield scanned the first human body part with magnetic resonance imaging (MRI); the finger of his assistant Andrew A. Maudsley [1]. MRI is an imaging technique, a non-invasive way of looking inside the human body. It can be used to visualise any part of the body, and is able to visualise certain tissues that are otherwise hard to visualise with other imaging techniques. One such example is the brain, which is why MRI is indispensable nowadays to study the brain in multiple sclerosis (MS). MRI was already mentioned in the discussion on what MS is (chapter 1) and how MS impacts cognitive performance (chapter 2). In this chapter, MRI is briefly described in the context of MS research and treatment. Figure 3.1 serves as a visual reference of the explanation.

3.1 How does MRI work?

3.1.1 Spinning protons

MRI is centred around the idea of a proton inside the nucleus of a hydrogen atom spinning around its axis (figure 3.1, panel 1). Because of this spin, a magnetic field is developed by the proton. Hence, we can think of the proton as a tiny magnet, characterised by a north and south pole [2].

3.1.2 Spinning out of control

The core idea of magnetic resonance imaging is that the spin of the protons can be aligned by subjecting them to an external magnetic field, and can subsequently be distorted by a radio-frequency (RF) pulse (figure 3.1, panel 2-3) emitted by an RF coil [3]. After this pulse, the proton relaxes to its

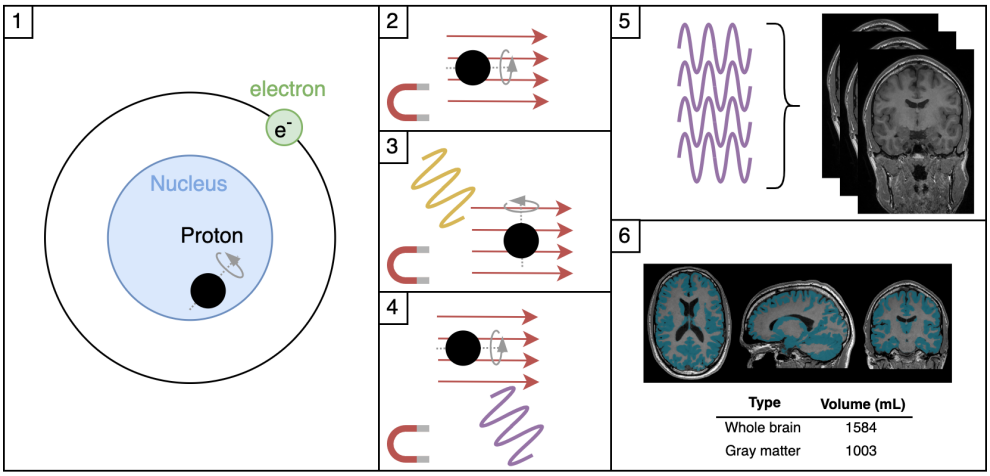

Figure 3.1: From proton spin to MR image and ultimately segmentation. The MR image in the illustration is a T1-weighted image.

equilibrium state (alignment with the external magnetic field), which causes a change in the magnetic flux in the RF coil, creating an electric current in the RF coil [3]. This current is the basis for the image to be constructed (figure 3.1, panel 4). Depending on the viewpoint on this relaxation, the relaxation is described with T1 (longitudinal) or T2 (transversal) relaxation. T1 is the time it takes to restore 63% of the longitudinal magnetisation (direction of the external magnetic field), while T2 is the time until 63% of the transverse magnetisation (plane perpendicular to the direction of the external magnetic field) is no longer present [4]. T2 is related to how fast the protons exchange the available energy [2].

3.2 Towards an image

By employing magnetic field gradients, spatial localisation can be encoded, thus an image can be created (figure 3.1, panel 5). This can for example be a 3D image to capture all dimensions of an anatomical structure. The 3D image is a cube, which in itself is divided into smaller cubes, volume units called "voxels". Voxels in a 3D image are analogous to pixels in a 2D image.

The intensity of each voxel in the image depends on the relaxation time, which by itself is determined by the contents of each voxel. For example, wa-

ter has a much longer T1 relaxation time compared to fat [2]. Since different tissues will have different compositions, they can be distinguished based on intensity.

The more voxels in an MR image of equal size (the same field of view), i.e. the smaller the voxels, the higher the resolution. This allows studying a body part in more detail; tissues from two different structures are less likely to be present in the same voxel. This has important implications for brain segmentation, which is discussed in the next section. In clinical practice, a resolution of 1mm isotropic (in all directions) is recommended for a 3D brain image [5], which will be the example used in this chapter to explain other concepts.

3.3 Brain segmentation

To visualise the brain, a 3D MR image of the head is made, usually including the upper cervical spinal cord. The total image easily contains millions of voxels, and allows a radiologist to spot image abnormalities. The radiologist will usually decide this empirically, based on extensive training including pattern recognition. For a quantitative description of image abnormalities, we however can resolve to brain segmentation (figure 3.1, panel 6).

Brain segmentation, as the word suggests, characterises certain "segments" in the brain. This can be a brain structure that is of particular interest, such as the hippocampus in Alzheimer's Disease (AD) [6] or the thalamus in studying cognitive impairment in MS [7]. Although this can yield important insights in brain anatomy, counting voxels would be tedious and time-consuming. Automated brain segmentation approaches have therefore been proposed in recent decades, for example using machine learning (ML, cfr. chapter 4) to identify structures. One example of an ML-based brain segmentation approach is the ico**brain** software of ico**metrix**, the industrial partner of this PhD thesis. This software was used for the study described in chapter 8 on predicting age from images. Technical details for this pipeline, or a newer version including deep learning (DL, cfr. chapter 4), can be consulted respectively in Jain et al. 2015 [8] and Rakić et al. 2021 [9]. An example of an ico**brain** T1 report (figure 3.3) and an ico**brain** ms report (figure 3.4 and figure 3.5) is included at the end of this chapter.

The typical result of a segmentation is a 3D map where each voxel contains

the probability of a voxel belonging to a class, which here is a certain brain structure. Voxels that exceed a certain probability threshold can then be counted and converted to a volume using the size of each voxel. In this way, a tabular representation of a brain image can be obtained of some segments of interest, which is commonly referred to as a "knowledge-based representation". Brain segmentation has been a key driver of biomarker research in MS, for example to study cognitive impairment (cfr. chapter 2). The value of MRI to study and treat MS are discussed next.

3.4 MRI for MS

The value of MRI in MS extends beyond diagnosis, which was discussed in chapter 1.

3.4.1 Clinical practice

In MS clinical routine, the recommended types of brain MR images depends on the context [5]. To assess disease activity and to monitor the efficacy of disease-modifying therapy (DMT, cfr. chapter 1), a T2-weighted and a T2 Fluid-Attenuated Inversion Recovery (FLAIR) image are recommended [5]. Both image types serve to better visualise inflammatory regions (termed lesions or plaques) in the brain. A T2 FLAIR image is a T2 image where, as the name suggests, the intensity of fluid is attenuated. Fluid will appear dark in the image, causing lesions to be distinguished more easily. Lesions appear hyper-intense on a FLAIR image, while they are hypo-intense on a T1 image (sometimes referred to as "black holes" [10]).

A contrast-enhanced T1-weighted brain image was recommended in the previous Magnetic Resonance Imaging in MS (MAGNIMS) guidelines [11] to detect new active lesions. Here, a T1 image is obtained as explained above, except a contrast is injected in the blood circulation of the patient prior to scanning. In a lesion, the blood-brain barrier can be disrupted, and contrast can exit the blood stream, enhancing the lesion (i.e. raising the intensity). This contrast agent typically contains gadolinium (Gd). Lastly, a T1-weighted brain image provides insights in brain volumetry, but is not necessary for routine follow-up of MS patients [5].

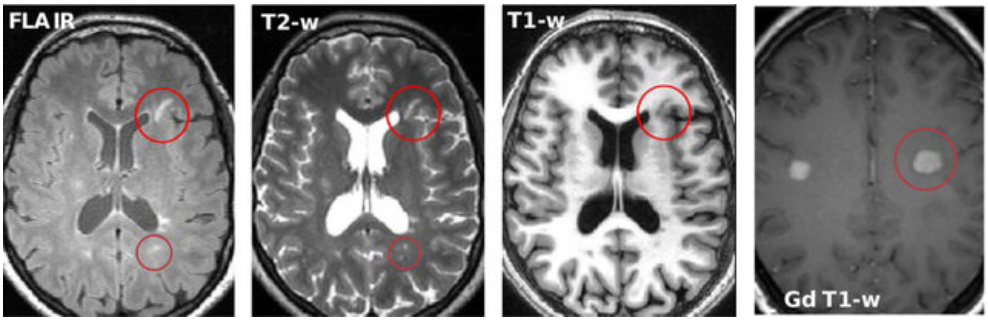

Figure 3.2: The most commonly used MR image types in MS. From left to right: FLAIR, T2-weighted, T1-weighted and Gadolinium-enhanced T1-weighted. The red circles highlight lesions, which are best visible in the FLAIR and Gd T1 image. Figure from Ma et al. 2022 [12] (unadapted), available under a CC BY-NC-SA 4.0 license (ResearchGate link: https://shorturl.at/cfp09).

3.4.2 Research

The image types used in clinical routine are also used for research purposes. After segmentation of the images, several imaging properties (such as lesion load, the total lesion volume) can be examined to establish image-derived biomarkers.

Another MRI technique that is used in MS research to identify biomarkers [13] is diffusion MRI. It visualises structural connectivity in the brain by focusing on the diffusion of water in the brain. A football player on a football pitch is free to move around in any direction, while a football player inside as soccer table game only can move in 1 direction. The same principle holds for water molecules in our body; they have more freedom to move outside cells, while inside cells, their movement is restricted [14]. Different tissues therefore have different diffusion properties, which can be visualised with MRI using a technique called diffusion weighted imaging (DWI). In the specific case of white matter in the brain, water molecules inside the axons of neurons move along the axon. By applying diffusion tensor imaging (DTI) to the diffusion weighted image, the white matter tracts can be reconstructed and white matter integrity assessed [14].

A popular MRI technique that is not used in clinical routine but for research purposes is functional MRI (fMRI). This technique relies on the blood-oxygen level dependent (BOLD) signal, reflecting the degree of blood oxygenation across the brain [15]. When more oxygen-rich blood is present in a

certain brain region, the brain region is thought to be more active; an indirect measure of activity. Contrasting earlier discussed techniques to visualise brain structure, fMRI measures brain function, which can then be used to map functional connectivity in the brain. However, the value of fMRI for analysing the MS brain was recently questioned based on MS-related data quality issues [16].

3.5 MRI for biomarker research

MRI has greatly advanced our understanding of biological underpinnings of MS symptoms, with biomarkers found in brain anatomy [17], as well as structural [18] and functional connectivity [19]. Although individual correlations are often reported, the relationship between brain imaging and MS symptoms remains paradoxical; brain damage as observed on MR images can be present without clinical repercussions and the other way round. This is commonly referred to as the clinico-radiological paradox [20].

How can this paradox be overcome? This is an ongoing investigation to which this thesis aims to contribute. Up until now, features that were extracted from images were knowledge-based; it is defined a priori which features to extract from an image, for example the grey matter volume of the brain. By converting the image to such representations, a lot of information is lost. Deep learning, an artificial intelligence (AI) technique explained in chapter 4, could however come up with new image representations by learning to map images to clinical symptoms. With this data-driven technique, all information is considered since it works with the original voxel space of the MR image. This representation is called the "latent space", and is a data-driven representation as no human knowledge was imposed. This latent space might provide new insights in the relationship between MRI and clinical symptoms at one cost; it is not understandable for humans. An interesting domain that aims to open this black box is called explainable AI (XAI), and is also discussed in chapter 4.

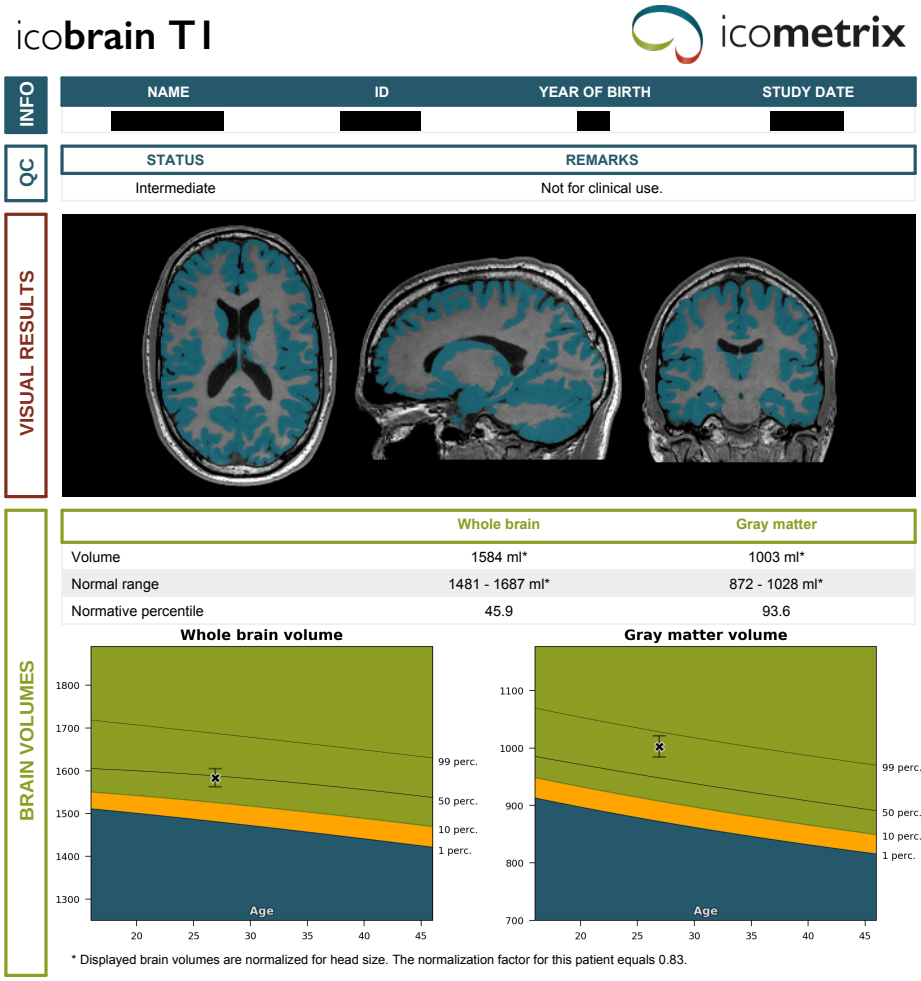

Figure 3.3: An ico**brain** T1 report (version 4.4.4). In the section "visual results", voxels with a high probability of being grey matter are coloured blue. Summing these voxels and correcting for intracranial volume allows comparing the volume with a healthy population as depicted in the section "brain volumes". ©2021 ico**metrix** NV, www.icometrix.com.

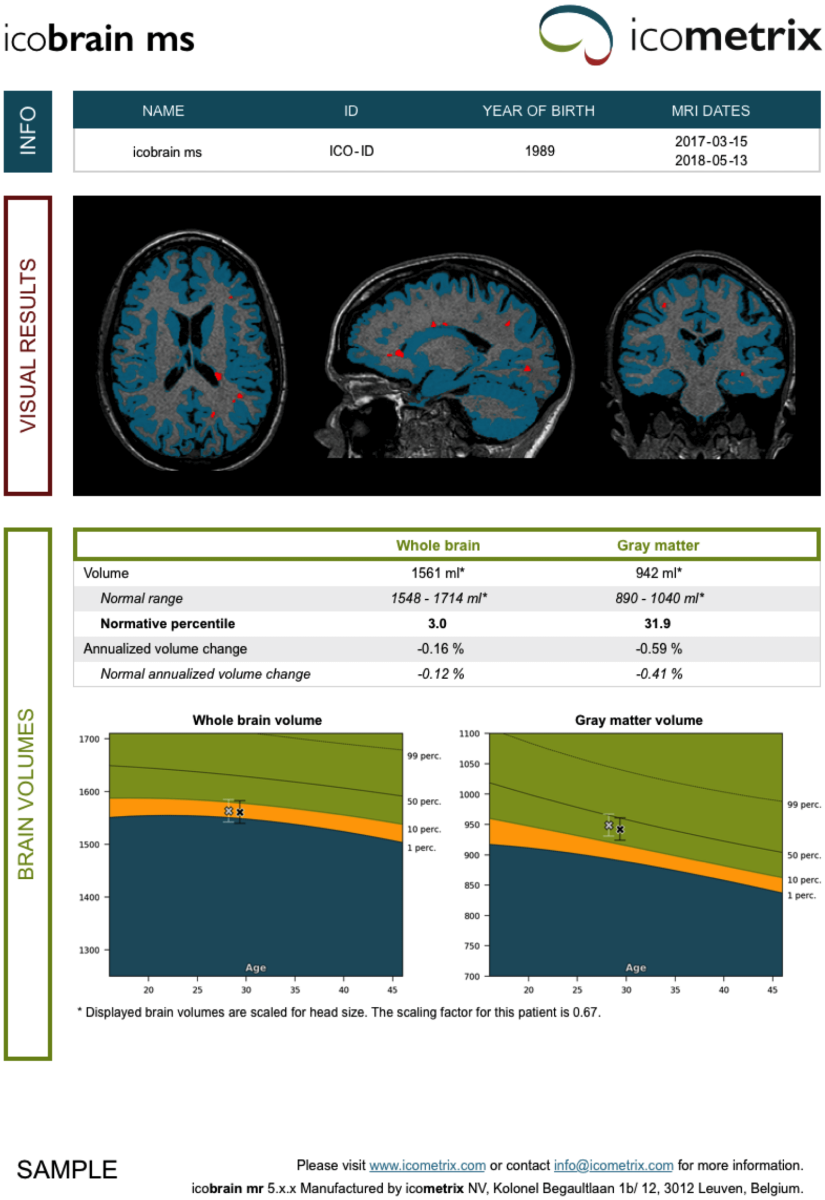

Figure 3.4: A sample ico**brain** ms report (page 1). This page is analogous to the ico**brain** T1 report in figure 3.3, except it also contains longitudinal information, i.e. the difference between two brain images. The voxels coloured red indicate loss of brain tissue. Image from https://icometrix.com/services/icobrain-ms, accessed on the 3rd of November 2023. ©ico**metrix** NV, www.icometrix.com.

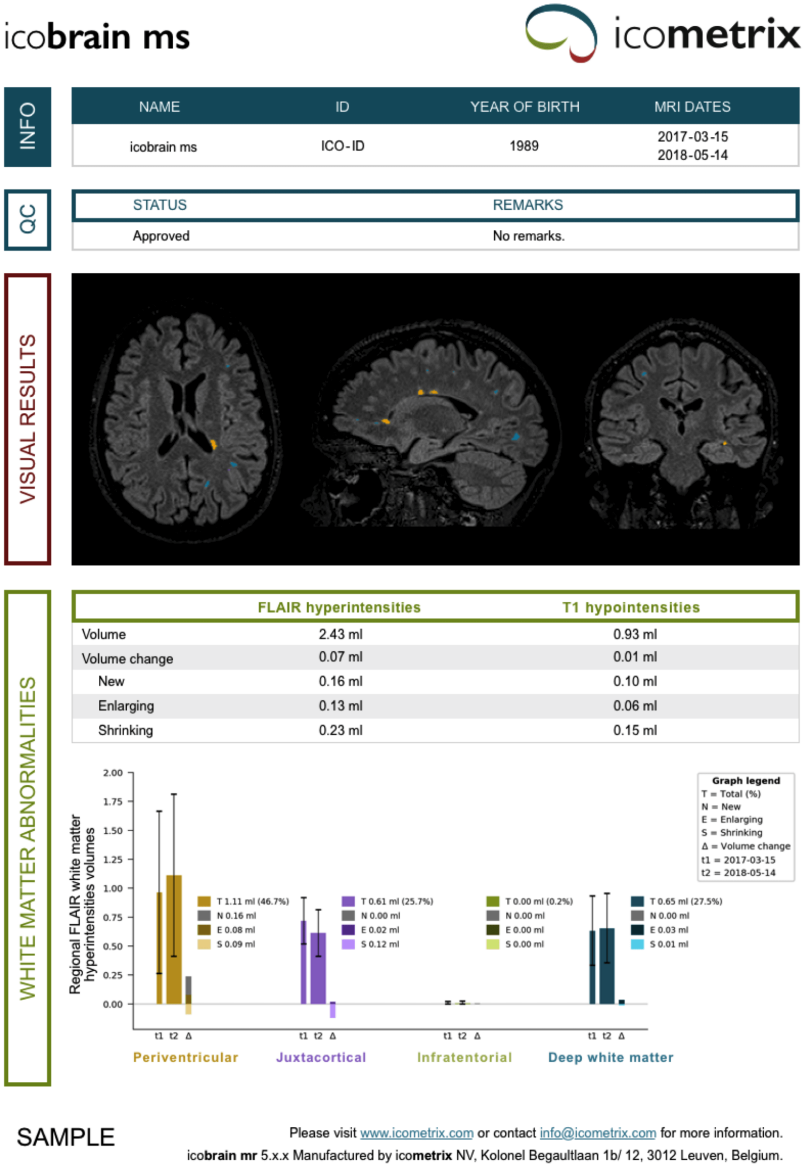

Figure 3.5: A sample ico**brain** ms report (page 2), indicating white matter lesions in 4 regions of an MS brain (colour coded). The table quantifies this information, both cross-sectionally (most recent MR image) and longitudinally (difference between 2 MR images). The latter is further categorised as new (N), enlarging (E) or shrinking (S). For the FLAIR column of the table, results are also presented graphically as bar charts. Image from https://icometrix.com/services/icobrain-ms, accessed on the 3rd of November 2023. ©ico**metrix** NV, www.icometrix.com.

References

[1] Mansfield, P. and Maudsley, A.A. Medical imaging by NMR. *The British journal of radiology*, 50(591):188–194, 1977.

[2] Grover, V.P., Tognarelli, J.M., Crossey, M.M., Cox, I.J., Taylor-Robinson, S.D. and McPhail, M.J. Magnetic resonance imaging: principles and techniques: lessons for clinicians. *Journal of clinical and experimental hepatology*, 5(3):246–255, 2015.

[3] Gruber, B., Froeling, M., Leiner, T. and Klomp, D.W. RF coils: A practical guide for nonphysicists. *Journal of magnetic resonance imaging*, 48(3):590–604, 2018.

[4] Pai, A., Shetty, R. and Chowdhury, Y.S. Magnetic resonance imaging physics. In *StatPearls [Internet]*. StatPearls Publishing, 2021.

[5] Wattjes, M.P., Ciccarelli, O., Reich, D.S., Banwell, B., de Stefano, N., Enzinger, C., Fazekas, F., Filippi, M., Frederiksen, J., Gasperini, C. et al. 2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *The Lancet Neurology*, 20(8):653–670, 2021.

[6] Van De Pol, L.A., Hensel, A., van der Flier, W.M., Visser, P.J., Pijnenburg, Y.A., Barkhof, F., Gertz, H.J. and Scheltens, P. Hippocampal atrophy on MRI in frontotemporal lobar degeneration and Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(4):439–442, 2006.

[7] Amin, M. and Ontaneda, D. Thalamic injury and cognition in multiple sclerosis. *Frontiers in Neurology*, 11:623914, 2021.

[8] Jain, S., Sima, D.M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., De Mey, J., Barkhof, F., Steenwijk, M.D., Daams, M. et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage: Clinical*, 8:367–375, 2015.

[9] Rakić, M., Vercruyssen, S., Van Eyndhoven, S., de la Rosa, E., Jain, S., Van Huffel, S., Maes, F., Smeets, D. and Sima, D.M. Icobrain ms 5.1: Combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions. *NeuroImage: Clinical*, 31:102707, 2021.

[10] Truyen, L., Van Waesberghe, J., Van Walderveen, M., Van Oosten, B., Polman, C., Hommes, O., Ader, H. and Barkhof, F. Accumulation of hypointense lesions (" black holes") on T1 spin-echo MRI correlates with disease progression in multiple sclerosis. *Neurology*, 47(6):1469–1476, 1996.

[11] Wattjes, M.P., Rovira, À., Miller, D., Yousry, T.A., Sormani, M.P., De Stefano, N., Tintore, M., Auger, C., Tur, C., Filippi, M. 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.

[12] Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., Cai, W., Barnett, M. and Wang, C. Multiple sclerosis lesion analysis in brain magnetic resonance images: techniques and clinical applications. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2680–2692, 2022.

[13] Baijot, J., Van Laethem, D., Denissen, S., Costers, L., Cambron, M., D'Haeseleer, M., D'hooghe, M.B., Vanbinst, A.M., De Mey, J., Nagels, G. et al. Radial diffusivity reflects general decline rather than specific cognitive deterioration in multiple sclerosis. *Scientific Reports*, 12(1):21771, 2022.

[14] Baliyan, V., Das, C.J., Sharma, R. and Gupta, A.K. Diffusion weighted imaging: technique and applications. *World journal of radiology*, 8(9):785, 2016.

[15] Hillman, E.M. Coupling mechanism and significance of the BOLD signal: a status report. *Annual review of neuroscience*, 37:161–181, 2014.

[16] Baijot, J., Denissen, S., Costers, L., Gielen, J., Cambron, M., D'Haeseleer, M., D'hooghe, M.B., Vanbinst, A.M., De Mey, J., Nagels, G. et al. Signal quality as Achilles' heel of graph theory in functional magnetic resonance imaging in multiple sclerosis. *Scientific Reports*, 11(1):7376, 2021.

[17] Matthews, P.M., Gupta, D., Mittal, D., Bai, W., Scalfari, A., Pollock, K.G., Sharma, V. and Hill, N. The association between brain volume loss and disability in multiple sclerosis: A systematic review. *Multiple Sclerosis and Related Disorders*, page 104714, 2023.

[18] Sbardella, E., Tona, F., Petsas, N., Pantano, P. et al. DTI measurements in multiple sclerosis: evaluation of brain damage and clinical implications. *Multiple sclerosis international*, 2013, 2013.

[19] Tahedl, M., Levine, S.M., Greenlee, M.W., Weissert, R. and Schwarzbach, J.V. Functional connectivity in multiple sclerosis: recent findings and future directions. *Frontiers in neurology*, 9:828, 2018.

[20] Barkhof, F. The clinico-radiological paradox in multiple sclerosis revisited. *Current opinion in neurology*, 15(3):239–245, 2002.

Chapter 4

Artificial Intelligence

4.1 What is AI?

A popular definition of artificial intelligence (AI) was introduced by Marvin Minsky in 1968, who was a pioneer in the field: "the science of making machines do things that would require intelligence if done by men" [1]. Two important questions can be raised related to this definition: when is a machine intelligent, and how can this intelligence be acquired?

4.2 When is a machine intelligent?

Alan M. Turing, who is regarded as one of the founders of AI, addressed the question "can machines think" in his paper "Computing Machinery and Intelligence" [2]. He did so by proposing the "imitation game", which is currently commonly known as the "Turing test".

The game-like situation consists of an interrogator who is in a room separate from a man and a woman. The objective of the interrogator is to identify the man and the woman by asking questions. The objective of the man and woman is to fool the interrogator. The key thought experiment is whether the interrogator can still be fooled if the man were to be replaced by a machine.

The ability to fool the interrogator, i.e., passing the Turing test, can be regarded as an attempt to establish whether a machine is able to "think", or is "intelligent". However, is this both a necessary and sufficient condition to be regarded "intelligent"? Might machines that fail this test still be considered as such? And are machines that pass the test really intelligent? As AI models

nowadays closely mimic human behaviour, but might still be fooled when asking the right questions, it is important to use the right assessment tools to reliably assess performance of an AI model [3].

4.3 How can intelligence be acquired?

While making abstraction of whether a machine is in fact intelligent, this section addresses how intelligence can be acquired. Two key concepts are presented, which are typically used in the modelling domain of this PhD thesis.

4.3.1 Rule-based AI

A rule-based AI is defined by rules that are imposed to a system by domain experts. When designing a system to decide whether an animal is either a cat or a dog, experts might for example create a decision tree (cfr. figure 4.1). In this case, expert knowledge is encoded into a succession of rules. When followed for data obtained from an animal of unknown class, the model will yield a prediction of the most probable class the animal belongs to.

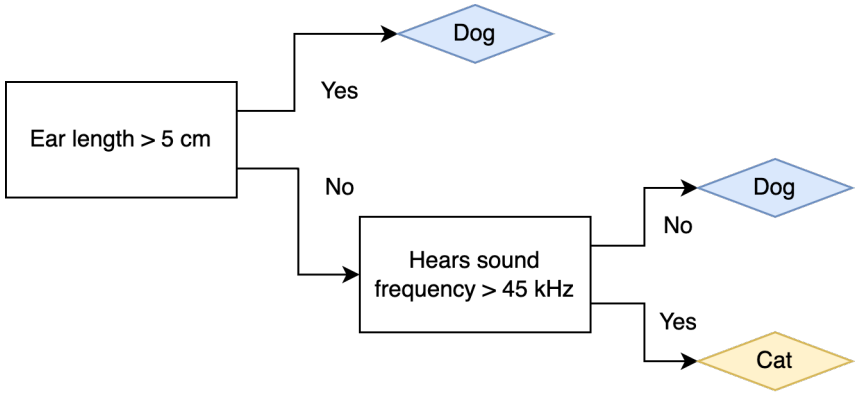

Figure 4.1: Simple example of a rule-based decision tree to classify cats and dogs.

4.3.2 Machine learning

Although the rule-based AI model is easy to understand, it is both rigid and difficult to establish; even profound domain expertise might not suffice to (1) find the variables that are key to a prediction, or (2) decide how to combine them in a model. Machine learning provides an alternative solution that might

tackle both problems.

In machine learning, a machine essentially learns from data. It can do so in three ways. In **supervised learning**, the machine is presented a certain input (for example an image of a cat or a dog) and the label that is associated to that input (the label "cat" or "dog"). The objective is to find a function that predicts the true label of a new input as accurately as possible. A more in-depth explanation of supervised learning by means of an example is added as appendix to this chapter. In **unsupervised learning**, the machine is only given the input data without any label. It is mostly used to find patterns in that data set, for example clustering observations. A third type is **reinforcement learning** which is for example intensively used in the gaming industry. Here, the computer learns by maximising a certain reward. In medical sciences, supervised learning is most commonly used.

In medical sciences, supervised learning is most commonly used. Let us consider the example of predicting the performance on a cognitive test from a structural MR image. The most intuitive approach, and the most logical continuation of prior research, is to use biomarkers of cognitive impairment (cfr. chapter 2) as input for the model. We can refer to this as a "knowledge-based representation"; domain expertise is used to transform the original input to a biomarker representation. We can then use more **traditional machine learning** algorithms to come up with a function, or model, that maps the input to the desired outcome label. An overview of some of such techniques is provided in the review paper [4] included in chapter 5.

The question might now arise why we do not simply use all available information, for example the original pixel space of an MR image. The catch here is the huge amount of data a computer must be able to process, and the complexity a model must be able to capture. Computational power nowadays allows training models that are capable of handling this complexity; a technique called **deep learning**. Deep learning is based on the idea illustrated in the appendix on supervised learning (cfr. figure 4.2). A neural network (panel 1) maps the input x to an output y. The input x is commonly referred to as the input layer, the weights w_1 and w_2 together form a hidden layer and y forms the output layer. In deep learning, the input layer usually contains much more information than a single scalar, for example a 3D image (3D tensor). The hidden layers are usually also more complex (more parallel nodes, i.e. the "width" of the network) and there are much more successive layers ("depth"

of the network), which together are capable of performing a highly complex mapping when stacked. The interesting property of a cascade of functions is that it is differentiable, and we can therefore optimise weights using the same technique as explained in figure 4.2 (stochastic gradient descent (SGD)), or a variation. When a network has been trained in this way, we can extract the output of the hidden layer immediately prior to the output layer, which we normally do not extract (hence the name "hidden"). This however is a representation that is analogous to the knowledge-based representation discussed earlier, except it is now "data-driven" and is termed the **latent space**. Altogether, this is referred to as the "feature extractor" [5]. The last mapping from this representation to the output label can essentially be seen as a traditional machine learning technique, except on highly abstract features. The specific deep learning technique covered in this thesis is a convolutional neural network (CNN), where some hidden layers (convolutional layers) essentially "filter" the image; they perform mathematical operations called "convolutions".

The choice of the techniques described above depend on the use case. Deep learning for example typically needs large datasets since it models a complex function on high-dimensional data. Another important consideration to make is the trade-off between accuracy and complexity that is currently given considerable attention in literature [6]. Rule-based approaches are interpretable but might fall short in solving a highly complex problem, while deep learning might be better to capture the complexity at the cost of interpretability; they are considered black box algorithms, meaning that the inner working of the algorithm is unknown. A new domain in AI has therefore emerged that is concerned with making these black box algorithms more transparent: explainable artificial intelligence (XAI) [7] .

4.4 Explainable AI (XAI)

The field of XAI is concerned with understanding AI models. It is a relatively young field of research, but is especially important in the real-world adoption of AI models in clinical practice, for example in terms of trust [8] and liability [9]. The need for XAI is high when models are complex, which is for example the case for deep learning. A comprehensive framework for XAI in the context of deep learning in medical imaging is presented in van der Velden et al. 2022 [10]. The framework categorises explanation methods based on three properties:

1. *The degree of enforcing inherent simplicity.* An explanation is model-based if the model is forced to be inherently simple enough to be un-

derstood (e.g. traditional machine learning methods), while if it is not enforced, the explanation occurs post hoc, i.e. after a model has been trained.

2. *How specific is the explanation method for a model type?* Model-specific versus model-agnostic, i.e. applicable to any type of model.

3. *The type of explanation.* Explaining the working of the entire model (global) or the output of a single case (local).

As listed in van der Velden et al. 2022, many XAI methods are available to better explain deep learning methods [10]. The suitability of an approach over the other is context-dependent, and should ideally be discussed with experts in the field. As they are the intended users of the models and have expertise in the domain, they are well-placed to guide XAI analyses, for example in terms of their specific needs to increase trust in the models.

One intuitive XAI method is for example occlusion sensitivity [11]. It is a post hoc method, model-agnostic, local explainability technique in which a part of the input image is "occluded" by a patch. The change in prediction error relative to the prediction error without occlusion is then calculated. By sliding the patch over the entire image, a "heatmap" is constructed indicating the performance drop per region when occluded. The severity of the performance drop indicates the relevance of the region for the prediction. When only occluding one voxel at a time, for a 3D MR image with millions of voxels, this can quickly become computationally very expensive. Using bigger patches reduces this at the cost of a lower resolution heat map with less precision to the locations of feature importance. A solution for this could be resolving to less computationally expensive methods, such as backpropagation-based methods [10].

4.5 Transfer learning

The concept of transfer learning dates back to the 1970s [12] with pioneering work of Bozinovski and Fulgosi [13]. In transfer learning, a model that is trained to perform a certain task (task A) is adapted to perform a related task (task B). This approach is especially interesting when there is few data or knowledge for task B, but an abundance for task A. Hence, a model can be reliably trained to perform task A, which can subsequently be fine-tuned to perform task B.

A relevant example in the domain of this thesis is fine-tuning a brain age model. Brain age has received significant interest in the last decade to study various neurological disorders [14]. It is a supervised machine learning technique, where a model is trained to predict the chronological age of a healthy person at the time of collecting certain brain information. This is usually a structural MR brain image, although other modalities are also used [15]. The predicted age is termed the "brain age", while subtracting chronological age from brain age is termed the "brain-predicted age difference" (BPAD) [16]. In various neurological disorders, including MS, this BPAD value is typically positive; brain age is overestimated, reflecting that the brain "looks older" than the person is at time of scanning [14]. As shown by Leonardsen et al. 2022 [17], brain age could be a good candidate for "task A" in transfer learning, as it can be reliably trained using tens of thousands of age-labelled MR images that are available in open source repositories. For multiple diseases, among which MS, their "task B" consisted of classifying people as either having or not having the disease.

4.6 Federated learning

In 2016, AI researchers from Google published a paper on training a model without sharing data; an approach they termed "federated learning". It is a paradigm shift in training machine learning models which is commonly done by first centralising all data. This is feasible for brain age research where data can easily be accessed, but becomes much harder when working with sensitive patient data. This data typically cannot leave the institution for privacy considerations or other legal constructions such as the general data protection regulation (GDPR). In federated learning, data remains at the original location and models are trained locally. The updates are then shared with a common server integrating the models, eventually yielding a global model that has been trained in a decentralised way.

4.7 AI in the context of MS

AI is primarily used for research purposes in MS. The overwhelming majority of research happens in the field of neurophysiology and radiology, for example to segment images (cfr. chapter 3), identify MS subtypes [18] or predict disease progression [4, 19]. However, especially segmentation approaches for MRI gradually find their way to clinical practice, such as the ico**brain** segmentation software discussed in chapter 3. Apart from those algorithms however,

the transition of AI models from research to the clinic is difficult. This might have numerous reasons, such as low performance, low explainability or other considerations (e.g. trust, fairness, bias). It has been hypothesised that the root problem of AI models not reaching clinical practice lies in the data, for example the quantity and quality [20]. Chapter 10 contains a discussion on the future role of AI in MS care.

This concludes the introductory section that provided the foundation for this PhD thesis. The next chapters concern the specific contribution of this PhD to the field, discussing three potential solutions for data scarcity in the domain of interest; exploring the link between structural MRI and cognition in MS using AI.

Appendix: Supervised machine learning example

Say that we observe that a certain input $\vec{x} = [5, 5]$, yields the output $y = 10$. We assume that this observation is the result of a linear equation of the form $y = \vec{x} \cdot \vec{w} = x_1 w_1 + x_2 w_2$ and want to find $\vec{w}$ that yields the closest mapping. The equation can be represented as a neural network with a certain input value x that is transformed by weights w_1 and w_2 to the output y (panel 1, figure 4.2).

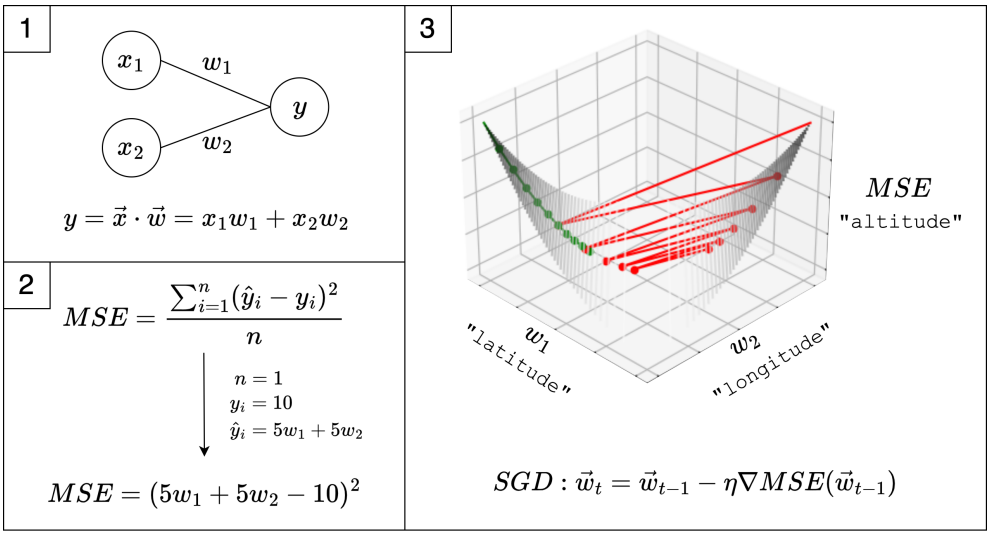

Figure 4.2: Minimising error with stochastic gradient descent (SGD)

To quantify how close the mapping is for a given set of $\vec{w}$, we need an error function. A popular error function is the mean squared error (MSE). Filling in the values for x and y in the MSE equation (panel 2) yields an equation plotted in grey in panel 3. The goal is to minimise this error function, for which we will use stochastic gradient descent (SGD). This will be explained with an analogy.

Let us think of the error function as 2 mountains separated by a valley. The x, y and z axis represent the latitude, longitude and altitude respectively. On each mountain, there is a giant. The goal of the giants is to approach the valley by each time taking a step in the steepest downwards going direction. Mathematically: each new position on time point t (w_t) is a step (η) taken from the old position (w_{t-1}) in the opposite direction ($-$ sign) of the steepest direction uphill ($\nabla MSE(\vec{w}_{t-1})$). The two giants have two different

approaches: one giant takes small steps (green), the other big steps (red). The former will gradually approach the lowest point of the valley, but rather slow. The latter overshoots; the giant steps over the valley. When running the simulation for 10 steps, the latter giant wins; the giant chose a better step magnitude (η, learning rate) and ended up at a coordinate ($\vec{w}$) of lower altitude.

References

[1] Stonier, T. The evolution of machine intelligence. In *Beyond Information: The Natural History of Intelligence*, pages 107–133. Springer, 1992.

[2] Turing, A.M. *Computing machinery and intelligence*. Springer, 2009.

[3] Biever, C. ChatGPT broke the Turing test-the race is on for new ways to assess AI. *Nature*, 619(7971):686–689, 2023.

[4] Denissen, S., Chén, O.Y., De Mey, J., De Vos, M., Van Schependom, J., Sima, D.M. and Nagels, G. Towards multimodal machine learning prediction of individual cognitive evolution in multiple sclerosis. *Journal of Personalized Medicine*, 11(12):1349, 2021.

[5] Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E. and Ganslandt, T. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

[6] Luo, Y., Tseng, H.H., Cui, S., Wei, L., Ten Haken, R.K. and El Naqa, I. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR| Open*, 1(1):20190021, 2019.

[7] Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[8] Diprose, W.K., Buist, N., Hua, N., Thurier, Q., Shand, G. and Robinson, R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, 27(4):592–600, 2020.

[9] Hacker, P., Krestel, R., Grundmann, S. and Naumann, F. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law*, 28:415–439, 2020.

[10] Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G. and Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

[11] Zeiler, M.D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference,*

Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pages 818–833. Springer, 2014.

[12] Bozinovski, S. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3), 2020.

[13] S. Bozinovski, A. Fulgosi (1976). The influence of pattern similarity and transfer of learning upon training of a base perceptron B2. (original in Croatian: Utjecaj slicnosti likova i transfera ucenja na obucavanje baznog perceptrona B2), Proc. Symp. Informatica 3-121-5, Bled.

[14] Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A. et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, 22(10):1617–1623, oct 2019.

[15] Engemann, D.A., Kozynets, O., Sabbagh, D., Lemaître, G., Varoquaux, G., Liem, F. and Gramfort, A. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *Elife*, 9:e54055, 2020.

[16] Cole PhD, J.H., Raffel MD, J., Friede PhD, T., Eshaghi MD, PhD, A., Brownlee PhD, FRACP, W.J., Chard MD, PhD, D., De Stefano MD, PhD, N., Enzinger MD, C., Pirpamer MSc, L., Filippi MD, FEAN, M. et al. Longitudinal Assessment of Multiple Sclerosis with the Brain-Age Paradigm. *Annals of Neurology*, 88(1):93–105, jul 2020.

[17] Leonardsen, E.H., Peng, H., Kaufmann, T., Agartz, I., Andreassen, O.A., Celius, E.G., Espeseth, T., Harbo, H.F., Høgestøl, E.A., de Lange, A.M. et al. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage*, 256:119210, 2022.

[18] Eshaghi, A., Young, A.L., Wijeratne, P.A., Prados, F., Arnold, D.L., Narayanan, S., Guttmann, C.R., Barkhof, F., Alexander, D.C., Thompson, A.J. et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nature communications*, 12(1):2078, 2021.

[19] Seccia, R., Romano, S., Salvetti, M., Crisanti, A., Palagi, L. and Grassi, F. Machine Learning Use for Prognostic Purposes in Multiple Sclerosis. *Life 2021, Vol. 11, Page 122*, 11(2):122, feb 2021.

[20] De Vos, M. and Van Schependom, J. Artificial intelligence will change MS care within the next 10 years: No. *Multiple Sclerosis Journal*, 28(14):2173–2174, 2022.

Chapter 5

Towards multimodal machine learning prediction of individual cognitive evolution in multiple sclerosis

Stijn Denissen[1,2], Oliver Y. Chén[3,4], Johan De Mey[1,5], Maarten De Vos[6,7], Jeroen Van Schependom[1,8], Diana Maria Sima[1,2†], Guy Nagels[1,2,9†]

1 AIMS Laboratory, Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, 1050 Brussels, Belgium **2** icometrix, 3012 Leuven, Belgium **3** Faculty of Social Sciences and Law, University of Bristol, Bristol BS8 1QU, UK **4** Department of Engineering, University of Oxford, Oxford OX1 3PJ, UK **5** Department of Radiology, UZ Brussel, Vrije Universiteit Brussel, 1090 Brussels, Belgium **6** Faculty of Engineering Science, KU Leuven, 3001 Leuven, Belgium **7** Faculty of Medicine, KU Leuven, 3001 Leuven, Belgium **8** Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, 1050 Brussels, Belgium **9** St Edmund Hall, Queen's Ln, Oxford OX1 4AR, UK

This chapter is based on a paper in the *Journal of Personalised Medicine* [1]

†These authors should be considered joint senior authors.

Abstract

Multiple sclerosis (MS) manifests heterogeneously among persons suffering from it, making its disease course highly challenging to predict. At present, prognosis mostly relies on biomarkers that are unable to predict disease course on an individual level. Machine learning is a promising technique, both in terms of its ability to combine multimodal data and through the capability of making personalised predictions. However, most investigations on machine learning for prognosis in MS were geared towards predicting physical deterioration, while cognitive deterioration, although prevalent and burdensome, remained largely overlooked. This review aims to boost the field of machine learning for cognitive prognosis in MS by means of an introduction to machine learning and its pitfalls, an overview of important elements for study design, and an overview of the current literature on cognitive prognosis in MS using machine learning. Furthermore, the review discusses new trends in the field of machine learning that might be adopted for future studies in the field.

Keywords
multiple sclerosis | prognosis | cognition | machine learning | artificial intelligence

5.1 Introduction

As one of the most puzzling neurodegenerative disorders, multiple sclerosis (MS) is characterised by a complex biological aetiology [2] and a highly heterogeneous disability progression. This gives rise to an important unmet need that has been given considerable attention in MS research in recent decades, which is the prediction of its future course [3, 4, 5, 6]. In light of an ongoing paradigm shift in medicine, moving from a disease-centred to a patient-centred approach [7], the ability to foresee disability build-up in a specific patient would be a true game changer in modern medicine; neurologists could intervene at an early stage, whereas patients and their caregivers could anticipate future challenges in daily life.

Currently however, to predict the natural course of MS on an individual level remains challenging. Foremost, the problem is intrinsically difficult since the disease manifests differently among patients. From a biological point of view, tissue damage in the central nervous system (CNS), caused by autoimmune processes, is not restricted to a single location or to a particular timepoint during the disease course [8]. Typical observations are the presence of lesions, resulting from processes such as demyelination and inflammation, in conjunction with the loss of CNS tissue [8]. However, MS patients typically present a wide range of clinical symptoms as well, ranging from motor and sensory impairments to fatigue, cognitive problems, and mental health issues [9]. Since every person with MS presents a unique biological and clinical profile, health-related predictions should be individualised.

At present, the best tools to estimate individual disease progression are the so-called prognostic biomarkers. They are defined by Ziemssen et al., 2019, as: "A prognostic biomarker" that "helps to indicate how a disease may develop in an individual when a disorder is already diagnosed" [10]. Although these variables can be regarded as the cobblestones of the road towards an accurate prognostic model, it is important to note that this term is assigned regardless of any magnitude of prognostic accuracy. Moreover, they are typically established at group level, which might be a suboptimal fit in light of the aforementioned heterogeneity across subjects with MS.

In a recent systematic review by Brown et al., 2020, the authors identified several studies that used various statistical techniques to combine prognostic biomarkers [3]. Although the techniques used are widespread, some studies report on the use of machine learning (ML), allowing personalised predictions

of the behaviour of a clinically relevant variable over time. The literature on this topic was synthesised by Seccia et al., 2021, although the authors limited their search to models using clinical data [5]. As can be expected from a young field of research, a sprawl of underlying methodology is observed among papers that use ML to perform prognostic modelling in MS; heterogeneity in terms of input features, learning algorithms, labels to predict, and assessment metrics hamper comparability among models. The narrative nature of both aforementioned reviews underscores the fact that quantitative synthesis by means of, e.g., meta-analysis or meta-regression, is not yet possible. Furthermore, various models aim to predict disease progression in terms of changes in the Expanded Disability Status Scale (EDSS), while a recent review by Weinstock-Guttmann et al., 2021, questions the use of the EDSS for prognostic purposes due to a lack of accuracy and stability [4]. This review also highlights the importance to look at other domains, such as cognitive impairment [4]. Problems in various cognitive domains are prevalent in persons with MS, especially in memory and information processing speed [11]. Since cognitive functioning was shown to be related to socio-economic aspects such as employment status [12] and income [13], prognostication in this domain could allow patients and their caregivers to anticipate future problems at an early stage.

Although the use of machine learning for cognitive prognosis is still in its infancy, this paper aims to offer directions in this field by (1) introducing the concept of machine learning, (2) outlining the pitfalls of machine learning in medical sciences, (3) offering guidance for the design of studies that use ML for cognitive prognosis using lessons learned from ML-powered physical prognosis, (4) summarising literature on ML-powered cognitive prognostication, and (5) highlighting trends in ML that could boost the field of MS prognosis. Since the main goal of this review is to provide directions for a young field of research rather than to synthesise the scarcely available literature, this review adopts a narrative, non-systematic design.

5.2 An Introduction to Machine Learning

Machine learning is defined in the Oxford University Press (OUP) as: “The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data” [14]. Although learning and adaptation can happen in multiple ways, typically categorised as “supervised”, “unsupervised” and “reinforcement” learning, the most com-

mon machine learning technique adopted in the medical sciences is supervised machine learning. The notion of "supervision" here is the presence of the ground-truth label to be predicted, which can either be a continuous variable (regression) or a categorical variable (classification). In general, the goal is to learn the relationship, in terms of a function, between a given input and the output—the ground-truth label. The function that subsequently best predicts the ground-truth label on input data that was not used to learn the function is the model of choice.

The concept can be clarified by means of an analogy; a student studying for a future exam. In the first phase, the student will gather knowledge on the domain by using available resources such as books and lecture notes (training). The student subsequently verifies whether additional study is necessary by completing an exam from previous years to which the answers are available (validation). Together, this is called the training phase. As necessary, training and validation are repeated until the student is ready to take the final exam, which constitutes the testing phase. Let us assume that we want to use supervised machine learning to predict a person's age given a brain magnetic resonance (MR) image. We start from a dataset with T1-weighted brain MR images (input) and the age at image acquisition (ground-truth label). Since age is a continuous variable, we are facing a regression problem. How we will learn the relationship between MRI and age depends on how we will use the MRI:

- Classical approach. The first approach is to analyse the brain MR images, yielding a set of features that describe the image such as volumetric quantifications of brain structures. This allows for the use of more classical supervised learning algorithms such as linear/logistic regression, support vector machines (SVM) and random forests (RF). The subsection below summarises some frequently used supervised learning algorithms;

- Deep learning. The second option is to use the raw brain MR images as input and use a technique called deep learning, which recently gained popularity as a subtype of machine learning. The major difference compared to classical machine learning is that it mitigates the necessity to manually transform raw data in a meaningful feature representation, the so-called "feature engineering" step, relying on human domain-specific knowledge [15]. Deep learning will automatically create meaningful representations from raw data, thus achieving representation learning [15]. This will typically yield "latent features", which are hard to interpret

by humans, but are deemed by the machine to be relevant. The advantage of deep learning lies in the more complex relationships that can be learned, while a major drawback is the need for large datasets, time and computational power.

5.2.1 Frequently used supervised learning algorithms

This section outlines supervised machine learning techniques exemplified for binary classification and univariate regression. For ease of interpretation, all examples use a low-dimensional feature space. However, the same principle holds when adding features towards higher-dimensional feature spaces.

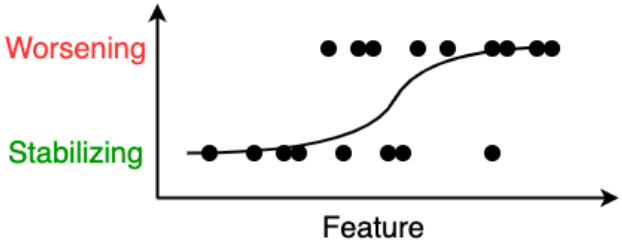

Figure 5.1: Schematic of logistic regression

Logistic regression. Logistic regression (figure 5.1)identifies the optimal sigmoid curve between the two labels to be predicted, yielding a probability of belonging to either of the two groups. In the illustration: the probability that a person will worsen or stabilise over time.

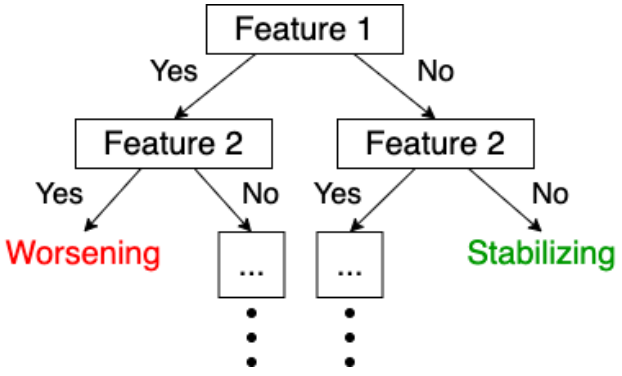

Figure 5.2: Schematic of a decision tree

Decision Tree. A decision tree (figure 5.2) is a sequence of decisions that are made on certain criteria. The last leaves of the tree indicate one of the class labels that are to be predicted.

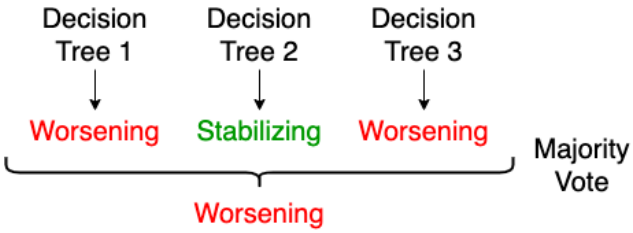

Figure 5.3: Schematic of a random forest

Random forest. This is an example of "ensemble learning", meaning that learning, and thus the resulting model, relies on multiple learning strategies, aiming to average the error out [16]. In this case, a random forest (figure 5.3) consists of multiple decision trees, mitigating the bias introduced by relying on one single decision tree. The ultimate prediction of a random forest classifier is the majority vote of the predictions of the individual decision trees in the random forest.

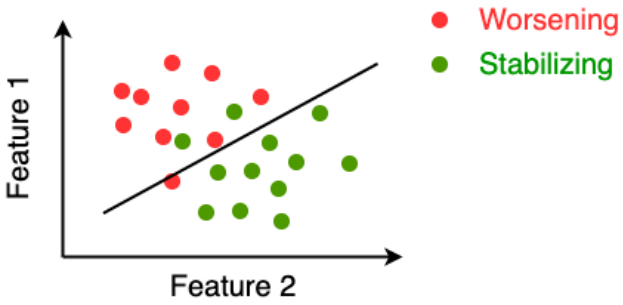

Figure 5.4: Schematic of a support vector machine (SVM)

Support vector machine. In case of two features, a support vector machine (SVM, figure 5.4) tries to find a line or a curve that separates the two classes of interest. It does so by maximising the distance between the line and the data-points on both sides of the line, thus maximally separating both classes.

Artificial Neural Network. An artificial neural network (ANN) was in-

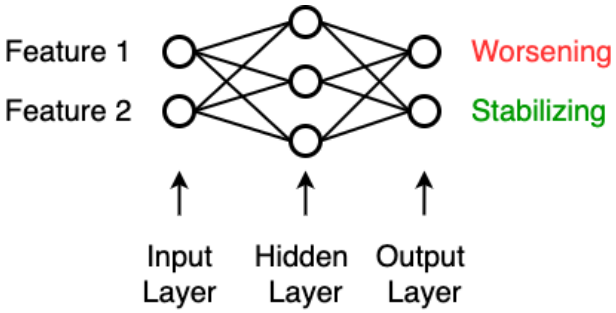

Figure 5.5: Schematic of an artificial neural network (ANN)

spired by the neural network of the brain and consists of nodes (weights) and edges that connect the nodes. Input data in either raw form or a feature representation enters the ANN on the left (input layer) and gets modified by the ANN in the hidden layers using the nodes' weights learned during the training phase, so that the input is optimally reshaped, or "mapped", to the endpoint that needs to be predicted on the right (output layer).

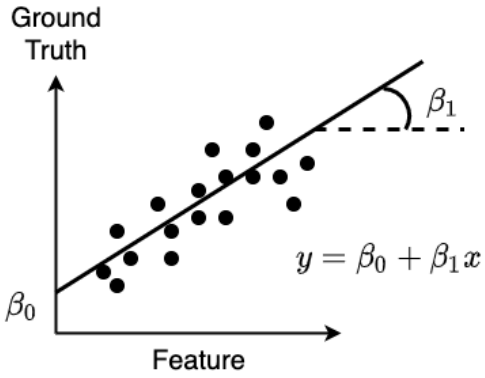

Figure 5.6: Schematic of linear regression

Linear Regression. Linear regression (figure 5.6is a technique in which the weight of every input feature is learned, which is multiplied with their respective feature and summed together with the so-called "bias" (also a learned weight but not associated to a feature, i.e., a constant), yielding a prediction that minimises the error with the ground-truth. In the 2D case, this is the line that minimises the sum of the squared vertical distances of individual points to the regression line. The learned weights in this case are the slope (β_1) and

intercept (β_0, bias) of the line.

5.3 Caveats for machine learning and potential solutions

Numerous pitfalls can be encountered when performing machine learning. The majority of them are generally applicable; they could arise in any machine learning query in any domain. Yet, we can encounter hazards that are specific for medical sciences. Both are discussed in this section, and solutions used in the field of prognostic modelling are also summarised.

5.3.1 General Pitfalls in Machine Learning

The most common pitfall in any machine learning query is overfitting. As already mentioned, a function is learned on training data and evaluated on validation and test data. Overfitting means that our learned function has become very specific to the training data, for example, because it also learned measurement errors in that dataset. Since measurement errors are different in another dataset, the function will be less accurate on that dataset. It is also possible, however, that we underestimate the complexity of the problem, which is the exact opposite case and understandably termed underfitting. For example, linear regression assumes a linear relationship between input features and the endpoint, which limits the model to only learn linear relationships, while the problem might be non-linear in reality. Figure 5.7 serves as a visual aid towards the understanding of under- and overfitting.

Overfitting often results from an imbalance between the number of variables and observations in the dataset. As a rule of thumb in the field, the number of observations should be at least 10 times as high as the number of variables [18]. To get to that ratio, we can address an imbalance in two ways: upscaling the observations or downscaling the variables. We note that in the case of downscaling variables, one should always remain vigilant not to underfit; informative features might be rejected as well.

1. Addressing the observations. Upscaling the number of observations is one way of tackling overfitting, but researchers often possess a database with a fixed number of observations. Nonetheless, several techniques exist to increase the number of observations based on those already present, e.g., using data augmentation. Although numerous variants exist, an

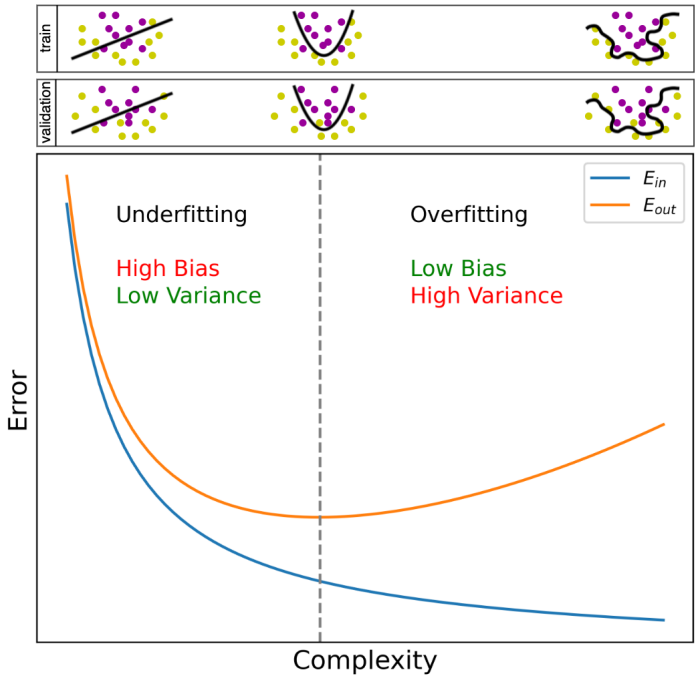

Figure 5.7: Bias–variance trade-off curve. Bias and variance vary according to model complexity [17]. The blue curve is E_{in}, the within-sample error representing the error on the training dataset. The more complex a function is allowed to be, the more specific the function becomes for the training dataset, i.e., overfitting. The latter is notable by the inception of an increase in E_{out} (orange curve, minimal value indicated with the vertical dotted line), the out-of-sample error, representing the error on the validation dataset. A simple function suffers high bias, i.e., it is highly likely to assume a wrong underlying function, since it only allows limited complexity between input and output to be learned (underfitting). By allowing more complexity, the bias decreases, but the function becomes highly variable depending on the dataset used for training (overfitting). An illustration is provided above, where the learned function is the line or curve separating two classes. From visual inspection, the optimal situation would be a smooth curve between the two classes (example in the middle). In the example on the left, underfitting occurs since only a straight line is allowed; many misclassifications occur in both training and validation data. In the example on the right, we observe a curve that squirms around all data points to fit the training dataset (overfitting), which, for example, happens when we allow the model to learn a complex function capable of learning measurement errors in a dataset. Hence, the function becomes specific to the training dataset; no misclassifications occur in the training data, but the same curve separating the validation dataset yields many misclassifications.

easy-to-grasp data augmentation method is the insertion of random noise into an observation [19], and can be interpreted as a similar, yet different subject record. A generative adversarial network (GAN) [20] serves the same purpose, which we will explain by means of a metaphor. Imagine a game-like situation in which a radiologist has to find out whether an image is a true MR image of the brain or was produced by a computer, i.e., a "villain", trying to fool the radiologist. Initially, the radiologist will easily identify which images were produced by the villain, since it had no clue how to generate a representative image. However, since the villain receives feedback on its effort, it will gradually start to understand how to create an image that will give the radiologist a hard time in telling whether it is a true image or a fake one. The radiologist on the other hand is forced to keep on improving classification skills, since it gradually becomes harder to distinguish them, in turn stimulating the villain to propose better images. Hence, the radiologist and the villain will infinitely stimulate each other to perform better. Ultimately, MR images are produced by the villain that could in fact have been the true ones, and which can subsequently be used to expand a dataset. Similar to deep learning, a GAN needs, besides time and computational power, large amounts and diversity of data to create qualitative new observations;

2. Addressing the features. The second option to restore an imbalance is the reduction of the number of features that the algorithm will be trained on. In feature selection, only the features that are deemed informative are selected. For an outline of several feature selection techniques in the context of medical sciences, we refer to Remeseiro et al., 2019 [21]. The original set of features can also be transformed to a new set of features. This can, for example, be done with principal component analysis (PCA), where we could say that the features are "reordered"; an equal number of features are obtained—the "principal components"—that explain variance in the data in a decreasing order. Feature selection can then occur on principal components instead of the original features. The additional benefit of PCA is that it is a solution to the problem of multicollinearity, in which features are mutually correlated. As a result, two variables might contain similar information, while the resulting principal components from PCA are uncorrelated [22].

Besides addressing observations and features, we discuss one additional technique to mitigate overfitting, which is training interruption. In their ef-

forts to predict the progression of disease, Bejarano et al., 2011 [23] and Yoo et al., 2016 [24] stopped the training phase early by monitoring the error in the validation set. As can be seen in figure 5.7, the error in the training set keeps reducing over time, since this is the goal of training. Initially, the same is observed for the validation data set, but upon obtaining a minimal value, the error will gradually increase, indicating the inception of overfitting. When stopping training at this point, overfitting might be mitigated.

Class imbalance is a specific pitfall for classification problems and is present when a certain class is overrepresented in the data, i.e., it contains more observations compared to the other class(es). For prognosis, subjects that do not worsen over time are often in the majority compared to worsening subjects [25, 26]. Like overfitting, it can lead to the poor generalisation of an algorithm [27]. Methods to correct class imbalance in a deep learning context are summarised in a systematic review by Buda et al., 2018 [27]. Two types of corrections are discussed, addressing either the data or the classifier itself. When addressing the data, we could restore the balance in two ways: by oversampling the minority class or by undersampling the majority class. On the other hand, we can make adjustments when training or testing the classifier. For example, one could decide to more severely penalise a misclassification towards a certain class compared to a misclassification towards another class, i.e., cost-sensitive learning [27]. These three methods were already explored in the light of prognostic modelling in MS to address the imbalance between stabilising and worsening subjects [25, 26].

5.3.2 Specific Pitfalls for Medical Data

Next to several general pitfalls, there are additional pitfalls when working with medical data:

1. Study data versus real-world data. Although the standardisation of conditions and minimising missing values are in general considered good practice, for example, when collecting data as part of a research study, it might limit the use of models in daily clinical routine that are known to be contaminated with, e.g., measurement errors, non-standardised test intervals, and missing values. When an algorithm encounters such inconsistencies during training, it could be expected to perform better on out-of-sample data. Although well-curated study data still dominate the field of prognostic modelling, efforts are underway to expand the use of real-world data [28, 29];

2. Single-centre versus multi-centre data. This argument is similar to the former; data from different clinical centres might be different due to discrepancies in testing equipment (e.g., MRI scanner), testing protocols, and patient characteristics. Introducing this heterogeneity already during the training phase might increase generalisability;

3. Multiple visits of the same patient. Finally, when using multiple visits of a patient as separate observations in a dataset, one should always remain vigilant that the visits do not get intermingled between train, validation, and test datasets. Since visits are often highly comparable, the performance on an unseen test dataset could be biased, performing better than would be the case when adopting a truly independent test dataset. This could be categorised under the hazard called "leakage", in which information of the test dataset leaks in the training dataset. To prevent this from occurring, Seccia et al., 2020 applied a correctional method called "leave one group out" (LOGO) [28]. With this method, they withdrew all visits of one subject from the training set and used them as a test set, after which the procedure was repeated for all subjects. This hinders models to recognise patients within a dataset. Other methods were, for example, discussed in Tacchella et al., 2018 [30] and Yperman et al., 2020 [29].

5.4 Designing an ML Study for Cognitive Prognosis

Supervised machine learning is popular for its ability to provide personalised predictions on health parameters that clinicians are used to work with in routine practice. One of these use cases includes predictions on how a patient with a certain condition progresses over time (prognosis) [31]. In the following section, we will address relevant questions when designing a machine learning study for cognitive prognosis in MS, using a question and answer (Q&A) approach. Answers are mostly constructed using lessons learned from the literature on ML-powered physical prognosis in MS and the literature on cognitive prognostic biomarkers.

5.4.1 Which Outcome to Predict?

As mentioned before, the outcome (categorical versus continuous) will define the type of problem we are facing: classification versus regression. When looking at cognitive outcomes, the most commonly affected domains are information processing speed and memory [11]. According to Sumowski et al.,

2018, information processing speed is best assessed with the Symbol Digit Modalities Test (SDMT), whereas for memory, the brief Visuospatial Memory Test—Revised (BVMT-R), California Verbal Learning Test—Second Edition (CVLT-II), and Selective Reminding Test (SRT) are the most sensitive tests [32]. However, composite scores also exist to provide a more holistic view on the cognitive status of persons with MS, which are summarised in Oreja-Guevara et al., 2019 [33]. In order to predict a change in these variables, a regression approach could include prediction of a future z-normalised test score [34], which is often the raw test score corrected for age, sex, and education level [34, 35]. For classification, a popular categorisation is defining "stable" and "declining" subjects [36], although wording can differ. In Filippi et al., 2013, for example, the authors defined cognitive worsening as an increase in impaired tests in a cognitive test battery over time, where impairment was defined as having a z-normalised test score below two [37]. Colato et al., 2021 defined worsening as a 10% decline of the SDMT score over time [38]. We furthermore note that practice effects can occur in cognitive tests over time [39]. To correct for this, a "reliable change index" was used in Eijlers et al., 2018 [36] and Cacciaguerra et al., 2019 [40]. Lastly, up until now, outcomes were all objective measures of cognition, while subjective, or self-reported measures also receive attention as outcomes for MS prognosis [41].

5.4.2 Which Features to Take into Account?

To be able to predict a future change in the variable of interest, the input of the machine learning model should receive careful consideration. Except when modelling on raw input data, learning should occur on features that are deemed informative towards the outcome to be predicted. To this end, we can use prognostic biomarkers, which were intensively studied in recent decades. However, although evidence on cognitive prognostic biomarkers exists, comprehensive reviews on the topic were mainly made for physical deterioration. We refer to reviews that summarise prognostic biomarkers for different modalities; demographics [42], clinical information [42], CNS imaging [43, 44, 45, 46], molecular information [10], and neurophysiology [46]. Yet, there appears to be an overlap between physical and cognitive prognostic biomarkers. Although it is beyond the scope of this review to provide a summary of cognitive biomarkers, we refer to studies that identified cognitive prognostic biomarkers for different modalities such as demographics [36, 47, 48], clinical information [36, 47, 48], MRI [36, 47, 49], optical coherence tomography (OCT) [50], molecular information [51], and neurophysiology [52]. In analogy with the previous question on outcomes, subjective measures might also be informative

for the prediction of disease course, such as patient-reported outcomes (PRO) [53]. Specifically for cognitive prognosis, features such as subjective cognitive impairment [48] and perceived ability to concentrate [54] were found to be informative.

5.4.3 On Which Time-Frame Should Predictions Be Made?

The literature usually makes a distinction between short-term and long-term prognosis. No clear cut-off between them has been reported, and this most probably depends on the clinical query that is addressed. Short-term prognosis is by far the most intensively studied [24, 26, 29], while Zhao et al., 2017 presented a longer-term predictive model of 5 years [25]. Yperman et al., 2020 stated that their rationale for a 2-year timeframe was based on maximising the number of observations in the dataset [29]. Data availability is highly likely to hinder the field in performing longer-term predictions using machine learning, but studies investigating prognostic biomarkers for long-term disability already show promising results [37, 47].

5.4.4 Which Machine Learning Algorithm to Use?

Given the heterogeneity in methodology throughout the literature, it is too preliminary to make firm statements regarding the superiority of one algorithm over another when considering performance. However, a second consideration is model complexity; linear models could underfit data, but are easy to interpret and familiar for clinicians. As illustrated by Sidey-Gibbons et al., 2019 [55], algorithms capable of handling increased complexity are in general harder to understand. This is, for example, the case for (deep) neural networks, which are often regarded as black box models [55].

5.4.5 How to Assess a Machine Learning Model?

Classifications will typically yield a so-called confusion matrix. In the case of a dichotomous endpoint, the confusion matrix is a 2×2 matrix with one axis indicating the true group labels and the other axis the predicted group labels. An example using the labels “worsening” versus “stabilising” is illustrated in figure 5.8, along with the metrics that can be calculated from this matrix. The different metrics allow us to study model performance from different perspectives. When looking at the confusion matrix of figure 5.8, low sensitivity will leave worsening patients undetected, which causes neurologists to falsely assume that their patient is stabilising. Withholding treatment—while this is in fact justified—will potentially endanger the patient’s well-being. The opposite

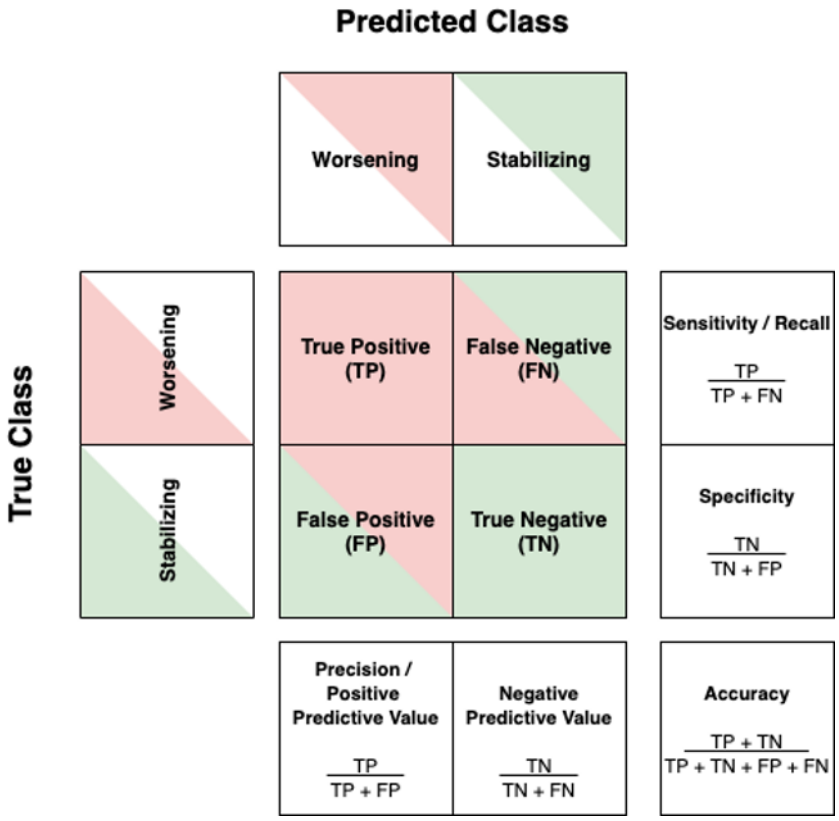

Figure 5.8: The confusion matrix and its derived metrics.

is true when we encounter low specificity; patients that do not worsen over time might receive treatment, while administration could potentially induce adverse events in their case.

Regarding regression performance, the most intuitive metric is the mean absolute error (MAE); it represents how much on average the predicted value deviates from the true value, while making abstraction of whether this is an under- or overestimation. The main difference with related metrics such as the normalised root-mean-square error (NRMSE, RMSE [56], MSE) is that MAE retains the unit of the outcome variable. Other performance metrics include the correlation between the true and predicted outcome [57], the variance explained by the input features (R^2) [56], and the Akaike Information Criterion [56].

5.4.6 How Should Authors Report the Performance of Their Machine Learning Model?

Solid interpretation and comparability of models stands or falls with how papers describe their methodology and performance. As discussed in the previous subsection, different performance metrics give different insights in model performance. Although the importance of a given metric mostly depends on the domain context, it is essential to not only report scores such as accuracy, sensitivity, and specificity, but also present the raw confusion matrix in classification problems. For regression, a 2-column data frame with the predicted and true ground-truth label allows the calculation of measures such as the MAE, NRMSE, RMSE, MSE, and correlation coefficient. Providing such results in publications (e.g., in supplementary materials [28]) would be a leap forward in terms of reproducible research, while the anonymity of subjects remains assured.

The benefit is twofold. Firstly, the readership of machine learning papers can extract other metrics that they are interested in. Secondly, it would also allow future reviews on machine learning models to move beyond a narrative design. In systematic reviews for example, meta-analysis and meta-regression allows for the quantitative synthesis of data, which is possible since randomised controlled trials (RCTs) are strongly recommended to adhere to the CONSORT statement [58], guiding RCT authors towards correct, transparent, and complete reports. Although the CONSORT statement is not applicable to machine learning research, another statement in the "Enhancing the QUAlity and Transparency Of health Research" (EQUATOR, https://www.equator-network.org/, accessed on 8 December 2021) network is in fact applicable: the "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis" (TRIPOD) statement [59].

5.4.7 When Is a Model Ready for Clinical Practice?

In order to introduce a predictive model in clinical practice, extensive technical validations and clinical performance evaluations are required, which should be complemented by ethical considerations and risk analysis. There needs to be maximal transparency towards the model's performance, so that regulators and clinicians can establish whether its error is acceptable in view of the potential risks to patients. However, when do we judge a machine learning model to be performant enough to be translated into a clinical decision support system (CDSS) [60]? In this regard, a first milestone is whether it performs better

than random, but in a second phase, it should compare favourably against other potentially simpler models, such as decision rules based on single prognostic biomarkers. Among other factors, model complexity might influence the trust of clinicians in artificial intelligence (AI) [61]. Furthermore, it would be informative to know how the machine's prognostic accuracy relates to the accuracy of human prediction, in this case of the neurologist. Although the literature on the latter is scarce, we identified one paper on the accuracy of decoding cognitive impairment in MS, albeit cross-sectional [62]. The authors found the accuracy to be comparable to chance, and highlighted the need for improved cognitive screening [62]. In order to benchmark how a model would perform in similar conditions to actual clinical practice, study designs should directly compare the prognostic accuracy of medical professionals without and with the assistance of the considered CDSS. A typical scenario involves comparing whether the CDSS helps bridging the gap between medical professionals with different levels of experience. For instance, an ongoing trial investigates prognostic accuracy of junior and senior doctors in the domain of traumatic brain injury [63].

We note that although some models might be complex, several methods exist to enhance clinicians' trust. In Tousignant et al., 2019 [26], deep learning, which is currently one of the most complex machine learning algorithms, was used to predict worsening in EDSS from MR images. The authors used a two-step process to gain the clinician's trust, namely by quantifying the model's confidence in its own predictions, and verifying whether the model is correct when it is confident [26]. We note that a whole field of research, i.e., explainable AI (XAI), is dedicated to, among other things, augmenting user trust [64].

5.4.8 Which Data to Use?

To address this question, we refer back to the section on "Specific Pitfalls for Medical Data", where we discussed study versus real-world data, single- versus multi-centre data, and dealing with multiple visits of the same subject.

5.5 State-of-the-Art ML-Powered Cognitive Prognostic Models

Literature in the field is scarce. This was confirmed by a PubMed search using the following search strategy: "(((multiple sclerosis[MeSH Terms]) OR

(multiple sclerosis)) AND ((cognit*) OR (cognition[MeSH Terms]))) AND ((((machine learning[MeSH Terms]) OR (machine learning)) OR (artificial intelligence[MeSH Terms])) OR (artificial intelligence))", which was run on 3 December 2021, and yielded 39 records. Among those, we identified two studies that used machine learning for cognitive prognosis; Kiiski et al., 2018 [57] and Lopez-Soley et al., 2021 [65]. Kiiski et al., 2018 used supervised machine learning on different combinations of multimodal data, including demographic, clinical, and electro-encephalography (EEG) data to predict short-term: (1) overall cognitive performance and (2) performance on information processing speed on a combined sample of persons with MS and healthy controls [57]. Lopez-Soley et al., 2021 also used multimodal data, including demographic, clinical, and MRI data, to predict short-term future cognitive impairment. This section is dedicated to the lessons that can be learned from their efforts.

5.5.1 Kiiski et al., 2018

First of all, the use of multimodal data is a good choice in light of the complex nature of MS and the identification of prognostic biomarkers in multiple domains. Moreover, the previous literature in the field of epilepsy established the superiority of multimodal data compared to using a single modality for machine learning predictions [66]. Secondly, the authors chose to z-normalise results for each neuropsychological test based on the mean and standard deviation (SD) of their sample, and use composite z-scores (average z-score of multiple tests) as the ground-truth label. A composite score was created for general cognitive functioning and one for information processing speed. Although transformation of raw test results allows comparison between, and aggregation of, different tests, the downside is in terms of clinical interpretation; clinicians have a reference frame for the original test results, whereas they do not for z-scores. Thirdly, the authors extracted over 1000 spatiotemporal features, whereas only 78 observations were used. This can be considered a large imbalance with a risk for overfitting, especially when considering the aforementioned rule of thumb of at least 10 times as many observations as features. The risk for overfitting might however have been reduced for several reasons:

- Using the "Elastic Net" [67] as learning algorithm. This is in essence a linear regression approach, but it uses regularisation, which is the addition of constraints to the learning process to increase a model's generalisability. Specifically, it uses a combination of L1 (Lasso) and L2

(Ridge) regularisation, which both tend to shrink large feature weights, whereas Lasso additionally tends to remove unimportant features from the model [67]. The low complexity of linear regression combined with regularisation might have increased generalisability;

- Using cross-validation (CV), which is a technique that allows the use of data for both training and validation purposes by training multiple models. If no CV were used, only one model would have been created on a part of the data, whereas validation would happen on the remaining data. Since this is a balance between few data for training (risk for a poorly trained model) and few data for validation (risk for a poor evaluation of the model), CV is a useful technique to minimise both risks.

Fourthly, we previously mentioned the importance of benchmarking to obtain a reference frame for the quality of the prediction. For this, the authors created a "null model" by shuffling the ground-truth values across subjects before starting the learning phase. According to the authors, this provides an intuition in the "level of optimism inherent in the model" [57]. Lastly, we highlighted that using a combined sample of persons with MS and healthy controls increases sample size, but it obfuscates a clear interpretation of its value for prognosis in MS. Their best-performing model for general cognitive functioning included all available data modalities and yielded a mean cross-validated correlation of 0.44.

5.5.2 Lopez-Soley et al., 2021

Opposed to the regression approach of Kiiski et al., 2018 [57], Lopez-Soley et al., 2021 used a classification approach to predict future global- and domain-specific cognitive impairment [65]. The risk of overfitting was reduced by using Lasso regularisation during logistic regression, 10-fold cross-validation, and retaining as much data as possible by imputing missing values.

Since cognitively impaired subjects were underrepresented for global cognition and all cognitive domains, it can be considered good practice that the authors used the "balanced accuracy" ((sensitivity + specificity)/2) to assess model performance across cognitive domains. The difference with accuracy (cfr. figure 5.8) can be clarified with an example. Say that in a dataset, 20 persons with MS experience cognitive decline, and 80 do not. If the model correctly classifies 70 of the 80 stabilising subjects, but only 5 of the 20 worsening subjects, the model achieves an accuracy of $(70 + 5)/100 = 75\%$. The

balanced accuracy, however is ((5/20) + (70/80))/2 = 56.25%. Hence, the balanced accuracy might be adopted for datasets that are unbalanced. For perfectly balanced datasets, accuracy and balanced accuracy yield the same value. Based on this metric, the authors reported the best performances for verbal memory (79%) and for attention/information processing speed (73%).

We note that by reporting the true class distribution, the authors greatly contributed to the interpretation of their result, as any evaluation metric can now be assessed with respect to that reference frame.

Overall, both studies yielded valuable intuition in the future design of machine learning studies for cognitive prognosis in MS. Despite the fact that predictions were obtained on a sample of both persons with MS and healthy controls in Kiiski et al., 2018 [57], predictive performances of both studies might serve as benchmarks for evaluating future studies in the field.

5.6 ML Trends and Opportunities for Prognostic Modelling in MS

Although studies dealing with prognostic modelling of cognitive evolution in MS are scarce, we see several interesting avenues for ML-driven prognostication in MS. We will discuss alternative approaches for prognostication, the simulation of treatment response and solutions to scarcity of longitudinal data.

5.6.1 Alternative Approaches for Prognostication

Hybrid predictions. Tacchella et al., 2018 introduced the proof-of-concept "hybrid predictions" [30] in the field of MS prognosis. The authors hypothesised that the discrepancy in "reasoning" between human and machine could in fact complement each other. Their results showed that the aggregation of human (medical students) and machine predictions consistently outperformed any of the single instances in predicting the conversion from relapsing–remitting to secondary progressive MS [30]. Besides performance, the fact that human intelligence is still involved in predictions could reassure clinicians that models do not solely rely on artificial intelligence, since they also rely on expert knowledge that algorithms might not be able to learn.

Digital twin. The field of machine learning for MS prognostication is mutually geared towards augmenting personalised care with personalised pre-

dictions. Since the prediction relies on the profile of a subject in terms of multimodal data, a subject can also be represented in a digital way, i.e., a digital twin. The concept of a digital twin was discussed elaborately in a recent review by Voigt et al., 2021, highlighting its potential to predict future disease course and simulate treatment effect [68].

5.6.2 Simulation of Treatment Response

Up until now, studies on prognostication mostly focused on predicting the natural course of multiple sclerosis. In our view, this is a necessary step to subsequently be able to predict, in a personalised way, how this natural course changes by administering certain treatment such as disease-modifying therapy (DMT). Although such estimates might be even more challenging, Pruenza et al., 2019 aimed to predict individual responses to 14 different DMTs [69]. The authors assigned a score per DMT that represented the likelihood of no disability progression in case of administration of the DMT [69]. Beyond a research effort, the authors created a tool that allows users to predict treatment response in new patients [69].

5.6.3 Solutions to Scarcity of Longitudinal Data

Transfer learning. A potential solution to scarcity of longitudinal data is to mitigate the necessity of building a model from scratch by using a robustly trained model from another domain, mostly related to the domain of interest. To this end, neural networks are typically used. Since the network's weights are meaningful to solve a related task, they could be used as initialisation for the task of interest, updating the weights using a smaller dataset. For example, Nanni et al., 2020 used pretrained networks (trained on the ImageNet database [70]) to classify pictures of everyday objects (number of pictures in the order of millions), for prognostic purposes in Alzheimer's disease (number of MR images in the order of hundreds) [71].

Federated learning. For various reasons, data sharing in medical sciences remains delicate [72], which might explain why efforts in ML-powered prognostication remain largely single-centre, extracting data from a single central database (centralised approach). However, an increasing number of studies [73, 74] prove that machine learning can also occur in a decentralised way, i.e., by federated learning, meaning that data remain at their original location, while still being used for machine learning in a remote location.

Continual learning. In continual learning, an AI is not trained once, but evolves over time by augmenting performance along with the ever-going supply of novel data. The implications of this technique in medical sciences are nicely discussed in Lee et al., 2020 [75].

5.7 Conclusions

Machine learning is a rising concept in light of clinical decision support systems and personalised medicine and could boost the quest to find a suitable predictive algorithm for prognosis in MS. Investigations should however also address cognitive deterioration, and authors should be maximally transparent in reporting their results to allow comparison in the field. In doing so, clinical decision support systems using machine learning to predict future cognitive deterioration in MS could become a reality in clinical practice, providing the best possible personalised care for persons with MS.

5.8 Key Messages

- Machine learning is capable of handling multimodal data and could predict disease course on an individual level;
- The literature on cognitive prognosis using machine learning in MS is scarce. Future studies on machine learning for prognosis in MS should not overlook cognitive deterioration;
- Recommendations for the design of studies on machine learning for cognitive prognosis are proposed;
- Researchers should aim to share as many results as possible to allow benchmarking, solid interpretation, and comparison in the field, for example, by sharing raw predictions;
- Several trends in machine learning could overcome current roadblocks in ML-powered prognostic modelling in MS, such as scarcity of longitudinal data.

References

[1] Denissen, S., Chén, O.Y., De Mey, J., De Vos, M., Van Schependom, J., Sima, D.M. and Nagels, G. Towards multimodal machine learning prediction of individual cognitive evolution in multiple sclerosis. *Journal of Personalized Medicine*, 11(12):1349, 2021.

[2] Winquist, R.J., Kwong, A., Ramachandran, R. and Jain, J. The complex etiology of multiple sclerosis. *Biochemical Pharmacology*, 74(9):1321–1329, nov 2007.

[3] Brown, F.S., Glasmacher, S.A., Kearns, P.K.A., MacDougall, N., Hunt, D., Connick, P. and Chandran, S. Systematic review of prediction models in relapsing remitting multiple sclerosis. *PLOS ONE*, 15(5):e0233575, may 2020.

[4] Weinstock-Guttman, B., Sormani, M.P. and Repovic, P. Predicting Long-term Disability in Multiple Sclerosis: A Narrative Review of Current Evidence and Future Directions. *International Journal of MS Care*, oct 2021.

[5] Seccia, R., Romano, S., Salvetti, M., Crisanti, A., Palagi, L. and Grassi, F. Machine Learning Use for Prognostic Purposes in Multiple Sclerosis. *Life 2021, Vol. 11, Page 122*, 11(2):122, feb 2021.

[6] Moazami, F., Lefevre-Utile, A., Papaloukas, C. and Soumelis, V. Machine Learning Approaches in Study of Multiple Sclerosis Disease Through Magnetic Resonance Images. *Frontiers in Immunology*, 0:3205, aug 2021.

[7] Lejbkowicz, I., Caspi, O. and Miller, A. Participatory medicine and patient empowerment towards personalized healthcare in multiple sclerosis. *Expert review of neurotherapeutics*, 12(3):343–52, mar 2012.

[8] Reich, D.S., Lucchinetti, C.F. and Calabresi, P.A. Multiple Sclerosis. *N Engl J Med*, 378(2):169–180, jan 2018.

[9] Kister, I., Bacon, T.E., Chamot, E., Salter, A.R., Cutter, G.R., Kalina, J.T. and Herbert, J. Natural History of Multiple Sclerosis Symptoms. *International Journal of MS Care*, 15(3):146, 2013.

[10] Ziemssen, T., Akgün, K. and Brück, W. Molecular biomarkers in multiple sclerosis. *Journal of Neuroinflammation*, 16(1):272, dec 2019.

[11] Macías Islas, M. and Ciampi, E. Assessment and Impact of Cognitive Impairment in Multiple Sclerosis: An Overview. *Biomedicines*, 7(1):22, mar 2019.

[12] Clemens, L. and Langdon, D. How does cognition relate to employment in multiple sclerosis? A systematic review. *Multiple sclerosis and related disorders*, 26:183–191, nov 2018.

[13] Kavaliunas, A., Karrenbauer, V.D., Gyllensten, H., Manouchehrinia, A., Glaser, A., Olsson, T., Alexanderson, K. and Hillert, J. Cognitive function is a major determinant of income among multiple sclerosis patients in Sweden acting independently from physical disability. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 25(1):104–112, jan 2019.

[14] Definition of Machine Learning. Oxford University Press. Available online: https://www.lexico.com/definition/machine_learning (accessed on 20 October 2021).

[15] Lecun, Y., Bengio, Y. and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

[16] Polikar, R. *Ensemble Machine Learning*. Springer US, Boston, MA, 2012.

[17] Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[18] Alibakshi, A. Strategies to develop robust neural network models: Prediction of flash point as a case study. *Analytica Chimica Acta*, 1026:69–76, oct 2018.

[19] DeVries, T. and Taylor, G.W. Dataset Augmentation in Feature Space. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, feb 2017.

[20] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144, jun 2014.

[21] Remeseiro, B. and Bolon-Canedo, V. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, 112(February):103375, 2019.

[22] Jolliffe, I.T. and Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, apr 2016.

[23] Bejarano, B., Bianco, M., Gonzalez-Moron, D., Sepulcre, J., Goñi, J., Arcocha, J., Soto, O., Carro, U.D., Comi, G., Leocani, L. et al. Computational classifiers for predicting the short-term course of Multiple sclerosis. *BMC Neurology*, 11:67, jun 2011.

[24] Yoo, Y., Tang, L.W., Brosch, T., Li, D.K.B., Metz, L., Traboulsee, A. and Tam, R. Deep Learning of Brain Lesion Patterns for Predicting Future Disease Activity in Patients with Early Symptoms of Multiple Sclerosis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10008 LNCS, pages 86–94. Springer Verlag, oct 2016.

[25] Zhao, Y., Healy, B.C., Rotstein, D., Guttmann, C.R., Bakshi, R., Weiner, H.L., Brodley, C.E. and Chitnis, T. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS ONE*, 12(4), apr 2017.

[26] Tousignant, A., Lemaître, P., Precup, D., Arnold, D.L. and Arbel, T. Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data. Technical report, 2019.

[27] Buda, M., Maki, A. and Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, oct 2018.

[28] Seccia, R., Gammelli, D., Dominici, F., Romano, S., Landi, A.C., Salvetti, M., Tacchella, A., Zaccaria, A., Crisanti, A., Grassi, F. et al. Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis. *PLOS ONE*, 15(3):e0230219, mar 2020.

[29] Yperman, J., Becker, T., Valkenborg, D., Popescu, V., Hellings, N., Wijmeersch, B.V. and Peeters, L.M. Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis. *BMC Neurology*, 20(1), mar 2020.

[30] Tacchella, A., Romano, S., Ferraldeschi, M., Salvetti, M., Zaccaria, A., Crisanti, A. and Grassi, F. Collaboration between a human group and

artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study. *F1000Research*, 6:2172, aug 2018.

[31] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, jan 2015.

[32] Sumowski, J.F., Benedict, R., Enzinger, C., Filippi, M., Geurts, J.J., Hamalainen, P., Hulst, H., Inglese, M., Leavitt, V.M., Rocca, M.A. et al. Cognition in multiple sclerosis: State of the field and priorities for the future. *Neurology*, 90(6):278–288, feb 2018.

[33] Oreja-Guevara, C., Ayuso Blanco, T., Brieva Ruiz, L., Hernández Pérez, M.Á., Meca-Lallana, V. and Ramió-Torrentà, L. Cognitive Dysfunctions and Assessments in Multiple Sclerosis. *Frontiers in Neurology*, 0(JUN):581, 2019.

[34] Ouellette, R., Bergendal, Å., Shams, S., Martola, J., Mainero, C., Kristoffersen Wiberg, M., Fredrikson, S. and Granberg, T. Lesion accumulation is predictive of long-term cognitive decline in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 21:110–116, apr 2018.

[35] Costers, L., Gielen, J., Eelen, P.L., Schependom, J.V., Laton, J., Remoortel, A.V., Vanzeir, E., Wijmeersch, B.V., Seeldrayers, P., Haelewyck, M.C. et al. Does including the full CVLT-II and BVMT-R improve BICAMS? Evidence from a Belgian (Dutch) validation study. *Multiple Sclerosis and Related Disorders*, 18:33–40, nov 2017.

[36] Eijlers, A.J.C., van Geest, Q., Dekker, I., Steenwijk, M.D., Meijer, K.A., Hulst, H.E., Barkhof, F., Uitdehaag, B.M.J., Schoonheim, M.M. and Geurts, J.J.G. Predicting cognitive decline in multiple sclerosis: a 5-year follow-up study. *Brain*, 141(9):2605–2618, jul 2018.

[37] Filippi, M., Preziosa, P., Copetti, M., Riccitelli, G., Horsfield, M.A., Martinelli, V., Comi, G. and Rocca, M.A. Gray matter damage predicts the accumulation of disability 13 years later in MS. *Neurology*, 81(20):1759–1767, nov 2013.

[38] Colato, E., Stutters, J., Tur, C., Narayanan, S., Arnold, D.L., Gandini Wheeler-Kingshott, C.A.M., Barkhof, F., Ciccarelli, O., Chard, D.T. and Eshaghi, A. Predicting disability progression and cognitive worsening

in multiple sclerosis using patterns of grey matter volumes. *Journal of neurology, neurosurgery, and psychiatry*, 92(9):995–1006, 2021.

[39] Portaccio, E., Goretti, B., Zipoli, V., Iudice, A., Pina, D.D., Malentacchi, G.M., Sabatini, S., Annunziata, P., Falcini, M., Mazzoni, M. et al. Reliability, practice effects, and change indices for Raos brief repeatable battery. *Multiple Sclerosis*, 16(5):611–617, may 2010.

[40] Cacciaguerra, L., Pagani, E., Mesaros, S., Dackovic, J., Dujmovic-Basuroski, I., Drulovic, J., Valsasina, P., Filippi, M. and Rocca, M.A. Dynamic volumetric changes of hippocampal subfields in clinically isolated syndrome patients: A 2-year MRI study. *Multiple Sclerosis Journal*, 25(9):1232–1242, aug 2019.

[41] Beier, M., Amtmann, D. and Ehde, D.M. Beyond depression: Predictors of self-reported cognitive function in adults living with MS. *Rehabilitation Psychology*, 60(3):254–262, aug 2015.

[42] Degenhardt, A., Ramagopalan, S.V., Scalfari, A. and Ebers, G.C. Clinical prognostic factors in multiple sclerosis: a natural history review. *Nature Reviews Neurology*, 5(12):672–682, dec 2009.

[43] Louapre, C., Bodini, B., Lubetzki, C., Freeman, L. and Stankoff, B. Imaging markers of multiple sclerosis prognosis. *Current Opinion in Neurology*, 30(3):231–236, jun 2017.

[44] Kearney, H., Miller, D.H. and Ciccarelli, O. Spinal cord MRI in multiple sclerosis—diagnostic, prognostic and clinical value. *Nature Reviews Neurology*, 11(6):327–338, jun 2015.

[45] Davda, N., Tallantyre, E., Neil, · . and Robertson, P. Early MRI predictors of prognosis in multiple sclerosis. *Journal of Neurology*, 266:3171–3173, 2019.

[46] Leocani, L., Rocca, M.A. and Comi, G. MRI and neurophysiological measures to predict course, disability and treatment response in multiple sclerosis. *Current Opinion in Neurology*, 29(3):243–253, jun 2016.

[47] Dekker, I., Eijlers, A.J.C., Popescu, V., Balk, L.J., Vrenken, H., Wattjes, M.P., Uitdehaag, B.M.J., Killestein, J., Geurts, J.J.G., Barkhof, F. et al. Predicting clinical progression in multiple sclerosis after 6 and 12 years. *European Journal of Neurology*, 26(6):893–902, jun 2019.

[48] Fuchs, T.A., Wojcik, C., Wilding, G.E., Pol, J., Dwyer, M.G., Weinstock-Guttman, B., Zivadinov, R. and Benedict, R.H. Trait Conscientiousness predicts rate of longitudinal SDMT decline in multiple sclerosis. *Multiple Sclerosis Journal*, 26(2):245–252, feb 2020.

[49] Hildesheim, F.E., Benedict, R.H.B., Zivadinov, R., Dwyer, M.G., Fuchs, T., Jakimovski, D., Weinstock-Guttman, B. and Bergsland, N. Nucleus basalis of Meynert damage and cognition in patients with multiple sclerosis. *Journal of Neurology*, pages 1–13, may 2021.

[50] Bsteh, G., Hegen, H., Teuchner, B., Amprosi, M., Berek, K., Ladstätter, F., Wurth, S., Auer, M., Di Pauli, F., Deisenhammer, F. et al. Peripapillary retinal nerve fibre layer as measured by optical coherence tomography is a prognostic biomarker not only for physical but also for cognitive disability progression in multiple sclerosis. *Multiple Sclerosis Journal*, 25(2):196–203, feb 2019.

[51] Gold, S.M., Raji, A., Huitinga, I., Wiedemann, K., Schulz, K.H. and Heesen, C. Hypothalamo–pituitary–adrenal axis activity predicts disease progression in multiple sclerosis. *Journal of Neuroimmunology*, 165(1-2):186–191, aug 2005.

[52] Nauta, I.M., Kulik, S.D., Breedt, L.C., Eijlers, A.J., Strijbis, E.M., Bertens, D., Tewarie, P., Hillebrand, A., Stam, C.J., Uitdehaag, B.M. et al. Functional brain network organization measured with magnetoencephalography predicts cognitive decline in multiple sclerosis. *Multiple Sclerosis Journal*, 27(11):1727–1737, oct 2021.

[53] Brichetto, G., Monti Bragadin, M., Fiorini, S., Battaglia, M.A., Konrad, G., Ponzio, M., Pedullà, L., Verri, A., Barla, A. and Tacchino, A. The hidden information in patient-reported outcomes and clinician-assessed outcomes: multiple sclerosis as a proof of concept of a machine learning approach. *Neurological Sciences*, 2019.

[54] de Groot, V., Beckerman, H., Uitdehaag, B.M., Hintzen, R.Q., Minneboo, A., Heymans, M.W., Lankhorst, G.J., Polman, C.H. and Bouter, L.M. Physical and Cognitive Functioning After 3 Years Can Be Predicted Using Information From the Diagnostic Process in Recently Diagnosed Multiple Sclerosis. *Archives of Physical Medicine and Rehabilitation*, 90(9):1478–1488, sep 2009.

[55] Sidey-Gibbons, J.A.M. and Sidey-Gibbons, C.J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19(1):64, dec 2019.

[56] Kuceyeski, A., Monohan, E., Morris, E., Fujimoto, K., Vargas, W. and Gauthier, S.A. Baseline biomarkers of connectome disruption and atrophy predict future processing speed in early multiple sclerosis. *NeuroImage: Clinical*, 19:417–424, 2018.

[57] Kiiski, H., Jollans, L., Donnchadha, S.Ó., Nolan, H., Lonergan, R., Kelly, S., O'Brien, M.C., Kinsella, K., Bramham, J., Burke, T. et al. Machine Learning EEG to Predict Cognitive Functioning and Processing Speed Over a 2-Year Period in Multiple Sclerosis Patients and Controls. *Brain Topography*, 31(3):346–363, may 2018.

[58] Schulz, K.F. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials. *Annals of Internal Medicine*, 152(11):726, jun 2010.

[59] Moons, K.G., Altman, D.G., Reitsma, J.B., Ioannidis, J.P., Macaskill, P., Steyerberg, E.W., Vickers, A.J., Ransohoff, D.F. and Collins, G.S. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*, 162(1):W1–W73, jan 2015.

[60] Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N. and Kroeker, K.I. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1), dec 2020.

[61] Asan, O., Bayrak, A.E. and Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22(6), jun 2020.

[62] Romero, K., Shammi, P. and Feinstein, A. Neurologists' accuracy in predicting cognitive impairment in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 4(4):291–295, jul 2015.

[63] Comparison of the Accuracy of the Neurological Prognosis at 6 Months of Traumatic Brain Injury Between Junior and Senior Doctors—Full Text View—ClinicalTrials.gov. Available online: https://clinicaltrials.gov/ct2/show/NCT04810039 (accessed on 2 November 2021).

[64] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G.Z. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), dec 2019.

[65] Lopez-Soley, E., Martinez-Heras, E., Andorra, M., Solanes, A., Radua, J., Montejo, C., Alba-Arbalat, S., Sola-Valls, N., Pulido-Valdeolivas, I., Sepulveda, M. et al. Dynamics and Predictors of Cognitive Impairment along the Disease Course in Multiple Sclerosis. *Journal of Personalized Medicine 2021, Vol. 11, Page 1107*, 11(11):1107, oct 2021.

[66] Memarian, N., Kim, S., Dewar, S., Engel, J., Jr. and Staba, R.J. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. *Computers in biology and medicine*, 64:67, sep 2015.

[67] Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.

[68] Voigt, I., Inojosa, H., Dillenseger, A., Haase, R., Akgün, K. and Ziemssen, T. Digital Twins for Multiple Sclerosis. *Frontiers in Immunology*, 0:1556, may 2021.

[69] Pruenza, C., Solano, M.T., Diaz, J., Arroyo-Gonzalez, R. and Izquierdo, G. Model for Prediction of Progression in Multiple Sclerosis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(6), 2019.

[70] Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. pages 248–255, mar 2010.

[71] Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., Castiglioni, I. and Initiative, T.A.D.N. Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. *Frontiers in Neurology*, 11:576194, nov 2020.

[72] van Panhuis, W.G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A.J., Heymann, D. and Burke, D.S. A systematic review of barriers to data sharing in public health. *BMC Public Health 2014 14:1*, 14(1):1–9, nov 2014.

[73] Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.C. and Shi, W. Federated learning of predictive models from federated Electronic

Health Records. *International Journal of Medical Informatics*, 112:59–67, apr 2018.

[74] Aledhari, M., Razzak, R., Parizi, R.M. and Saeed, F. Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE access : practical innovations, open solutions*, 8:140699–140725, 2020.

[75] Lee, C.S. and Lee, A.Y. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, jun 2020.

Part II

Three solutions for data scarcity

Chapter 6

Hypotheses

The previous chapters provided the foundation on which this PhD thesis is built. In synthesis, multiple sclerosis is an inflammatory (auto-immune) and neurodegenerative disease of the central nervous system that is typically diagnosed in people between 20 and 30 years old. The damage is visible by inflammatory plaques and neurodegeneration on MR images of the brain and spinal cord, and leading to a wide range of difficulties including cognitive impairment. The relationship between the observed damage and the clinical symptoms however remains poorly understood, known as the clinico-radiological paradox. Artificial intelligence has been proven to be applicable to brain MR images and therefore could shed new light on the paradox. Especially deep learning might come up with data-driven representations of the brain that could guide clinicians' way of examining MR images in the light of cognitive impairment.

Although the computer hardware nowadays allows to perform this deep learning research, this thesis aims to overcome the core limiting factor of deep learning research in a medical context: data availability. Data is scattered across clinical centres and data sharing is difficult due to privacy considerations; data is protected by the General Data Protection Regulation (GDPR). To still be able to perform deep learning research, this thesis presents three solutions: 1) facilitating the collection of data, 2) reducing the need for data and 3) increasing the accessibility of data. They form the four key hypotheses of the thesis.

Hypothesis 1: icognition, a smartphone-based cognitive screening battery, is valid and reliable for people with MS.

The first hypothesis relates to facilitating data collection for both care and research in MS. Chapter 7 introduces ico**gnition**, which is a smartphone-based cognitive screening battery that aims to assess the two most commonly impaired cognitive domains in MS: memory and information processing speed [1]. As most people nowadays have a smartphone, ico**gnition** could be performed at home, allowing more frequent cognitive follow-up and telemedicine (practising medicine remotely using technology). From an AI perspective, digitalising cognitive tests has the great benefit that data is stored digitally immediately, facilitating the construction of digital research databases. Chapter 7 contains the study that was carried out to assess the hypothesis that ico**gnition** is valid and reliable for people with MS.

Hypothesis 2: Brain age correlates with cognitive performance in people with MS

The concept of brain age was introduced in chapter 4. On average, the brain age of people with MS is higher than their chronological age, indicating that their brains "look older" [2]. We hypothesise that brain age itself, and the degree to which the age is overestimated, correlates with the performance of people on the symbol digit modalities test (SDMT), a measure of information processing speed [3].

Hypothesis 3: Transfer learning allows fine-tuning brain age models to predict cognition

The third hypothesis is related to reducing the need for data. As explained in chapter 4, a model that is trained for a certain task can be fine-tuned to carry out another task, given that the tasks are sufficiently similar. We hypothesise that this is the case for predicting age (task A) or cognition (task B) from a brain MRI. A brain age model is much more easy to train since it relies on healthy control data (brain MRI and age at scanning) that are abundantly available in open source repositories. With having access to thousands of images, deep neural networks can be robustly trained to predict age from brain MRI. In chapter 8, the condition for transfer learning is checked, i.e. the

similarity between predicting age and cognition from brain MRI. In chapter 9, the actual transfer learning is performed by fine-tuning the task-specific layers of a deep brain age network on a smaller data set of people with MS, including brain MRI and performance on a test for information processing speed (SDMT, chapter 2).

Hypothesis 4: Federated learning is feasible for deep learning research in multiple sclerosis

Traditional machine learning research is performed on centralised data sets. That is, data are located on a central server, which requires data to be shared among clinical centres. As explained in chapter 4, the concept of federated learning is to share models instead of data. This allows **decentralised** machine learning, meaning that models are trained locally, while data remain on the original location.

This hypothesis concerns proving the feasibility of performing this approach with international clinical MS centres. The aim is to establish the first federated learning network for cognitive neuroscience in MS, and prove the feasibility of training a model with this network (chapter 9). The model to be trained in a decentralised way is a model that predicts cognition from brain MRI, by using the transfer learning concept mentioned in hypothesis 2.

References

[1] Macías Islas, M.Á. and Ciampi, E. Assessment and impact of cognitive impairment in multiple sclerosis: an overview. *Biomedicines*, 7(1):22, 2019.

[2] Cole PhD, J.H., Raffel MD, J., Friede PhD, T., Eshaghi MD, PhD, A., Brownlee PhD, FRACP, W.J., Chard MD, PhD, D., De Stefano MD, PhD, N., Enzinger MD, C., Pirpamer MSc, L., Filippi MD, FEAN, M. et al. Longitudinal Assessment of Multiple Sclerosis with the Brain-Age Paradigm. *Annals of Neurology*, 88(1):93–105, jul 2020.

[3] Benedict, R.H., DeLuca, J., Phillips, G., LaRocca, N., Hudson, L.D. and Rudick, R. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. *Mult Scler*, 23(5):721–733, April 2017.

Chapter 7

icognition: a smartphone-based cognitive screening battery

Stijn Denissen[1,2†], Delphine Van Laethem[1,3†] Johan Baijot[1], Lars Costers[1,2], Annabel Descamps[2], Ann Van Remoortel[4], Annick Van Merhaegen-Wieleman[5], Marie B D'hooghe[4,6], Miguel D'Haeseleer[4,5,6], Dirk Smeets[2], Diana Maria Sima[2], Jeroen Van Schependom[1,7], Guy Nagels[1,2,5,8]

1 AIMS Lab, Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, Brussel, Belgium **2** icometrix, Kolonel Begaultlaan 1b, 3012 Leuven, Belgium **3** Department of Physical and Rehabilitation Medicine, UZ Brussel, Brussel, Belgium **4** Neurology Department, National Multiple Sclerosis Center, Melsbroek, Belgium **5** Neurology Department, UZ Brussel, Brussel, Belgium **6** Center for Neurosciences, Vrije Universiteit Brussel, Brussel, Belgium **7** Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussel, Belgium **8** St Edmund Hall, University of Oxford, Oxford, UK

† Denissen S. and Van Laethem D. should be considered joint first author.

Under review
This chapter is based on a preprint on *medRxiv* [1]

Abstract

Background: Cognitive deterioration is prevalent in multiple sclerosis (MS) and requires regular follow-up, which is time-consuming and costly. Telemedicine could offer a solution, as it is feasible and well-accepted by people with MS. Current smartphone-based applications however focus solely on information processing speed, while memory is also commonly affected.

Objectives: To validate a smartphone-based cognitive screening battery, ico**gnition**, to signal deterioration in both memory and information processing speed.

Methods: ico**gnition** consists of three tests (Symbol Test, Dot Test and visual Backwards Digit Span (vBDS)). These tests are based on validated paper-pencil tests: the Symbol Digit Modalities Test (SDMT), the 10/36 Spatial Recall Test (SPART) and the auditory Backwards Digit Span (aBDS), respectively. To establish the validity of ico**gnition**, 101 people with MS and 82 healthy subjects completed all tests. 21 healthy subjects repeated testing 2 to 3 weeks later.

Results: All tests in ico**gnition** correlate well with their paper-pencil equivalent (Symbol Test: r=.67, P<.001; Dot Test: r=.31, P=.002; vBDS: r=.69, P<.001), negatively correlate with the Expanded Disability Status Scale (EDSS: Symbol Test: ρ=-.34, P<.001; Dot Test: ρ=-.32, P=.003; vBDS: ρ=-.21, P=.04) and show moderate-to-good test-retest reliability (Symbol Test: ICC=.85, r=.85, P<.001; Dot Test: ICC=.73, r=.74, P<.001; vBDS: ICC=.81, r=.83, P<.001). Test performance was comparable between people with MS and healthy subjects for all cognitive tests, both in ico**gnition** and the gold standard paper-pencil tests.

Conclusion: ico**gnition** is a valid and reliable tool to remotely screen cognitive performance in persons with MS.

Keywords
multiple sclerosis | telemedicine | smartphone | cognition | memory | information processing speed

7.1 Introduction

Medicine is increasingly digitalising, and there are solid reasons to stimulate this trend. Doctors can more easily access and share electronic health records, and storing data in a digital format facilitates visualisation and organisation in research databases, yielding new insights into pathology and disease management. Beyond a nice-to-have, digital medicine was a crucial element in handling the COVID-19 pandemic, allowing telemedicine to be practiced when social distancing was necessary.

Telemedicine provides practical solutions for people with multiple sclerosis (MS). Telemedicine tools are well-accepted by patients, and their feasibility and cost-effectiveness have been established previously [2]. Patients moreover tend to objectively benefit from these approaches, for example for fatigue management [3] and improving cognitive function [4]. The latter is important as nearly half of the people with MS have cognitive impairment [5], which has significant repercussions on daily life activities and societal participation, employment and susceptibility to psychiatric disorders [6]. Digital solutions for assessing cognitive impairment are currently emerging to allow frequent screening.

There are however two limitations to current digital cognitive assessments. First, they focus primarily on information processing speed (IPS). Besides slowed IPS however, the hallmark cognitive problem in MS is impaired memory [7], which is not assessed by current smartphone-based cognitive assessments [8]. Although memory assessment does exist on a tablet [9], tablets are less suitable for consistent follow-up as they are used far less frequently compared to smartphones. Second, current smartphone applications might be prone to motor interference. They are predominantly digital versions of the symbol digit modalities test (SDMT) [10], which is a popular test in clinical practice to measure information processing speed and has excellent psychometric properties [11]. Digitalising the SDMT allows randomising its key, which could reduce practice effect as reported in Pereira et al. 2015 [12]. Creating an exact digital replica of the SDMT however requires patients to choose from nine small buttons on the screen, which could cause motor interference as fine motor skills are commonly affected in MS [13].

To tackle these limitations, we here validate a new smartphone-based cognitive screening battery called "ico**gnition**". It is a quick smartphone-based screening tool for remote follow-up of the two most commonly impaired cog-

nitive domains in MS; information processing speed and memory [7]. It is intended to be part of the recently established ico**mpanion** application, a digital diary for people with MS [14]. Regular remote screening could lead to a faster confirmation of cognitive deterioration by a neuropsychologist and addressing this deterioration by the patient's neurologist.

7.2 Methods

7.2.1 Participants

Inclusion criteria for people with MS were a confirmed diagnosis of multiple sclerosis according to the McDonald criteria [15]. MS subjects were excluded if they were hospitalised for reasons other than rehabilitation or if they had a relapse within the last month. Both persons with MS and healthy control subjects were excluded in the presence of any other neurological or psychiatric disorder or learning disorder. A total of 101 people with MS and 82 healthy control subjects (matched on age, sex, and education level) met the in- and exclusion criteria for this study. All subjects were either Dutch-speaking or bilingual including Dutch and were 18 years or older.

7.2.2 Ethics

This study was approved by the "Commissie Medische Ethiek" of the UZ Brussel (B.U.N. 143201940335) and the National MS Center of Melsbroek. All participants signed informed consent prior to inclusion.

7.2.3 icognition

The ico**gnition** cognitive screening battery consists of three tests (figure 7.1).

The Symbol Test is based on the computerised digit-symbol substitution test (DSST) presented in Rypma et al. 2006 [16]. A combination of symbols is presented to the subject, one at a time. Furthermore, a key is presented on top, which consists of 9 pairs of symbols, and which shuffles every trial. For every trial, the subject needs to indicate whether the combination occurs in the key on top. The total score is the number of correct answers in 90 seconds. This test is designed to capture information processing speed.

The Dot Test consists of three phases. In a first phase, a subject is presented a 4x4 grid in which three dots are shown for 3 seconds. Then, as a means

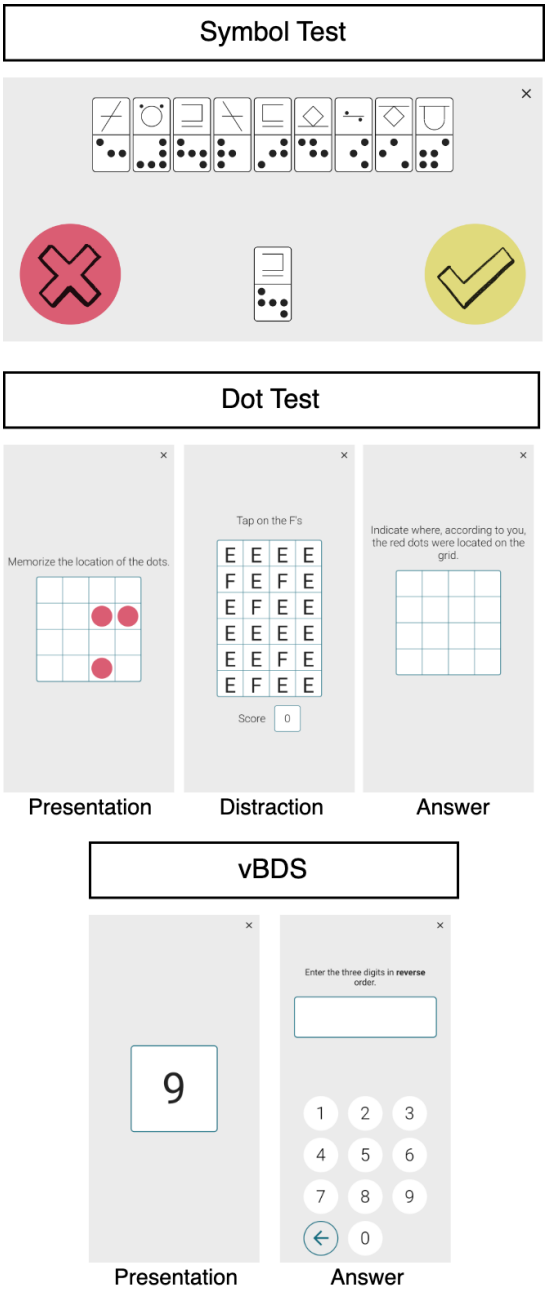

Figure 7.1: Screenshots of the ico**gnition** tests. Note: although the instructions were in Dutch during testing, they are presented here in English. vBDS = visual Backwards Digit Span.

of distraction, the subject is presented a 4x6 grid of "E" and "F" shapes and has to identify as many "F" shapes as possible in 4 seconds. In the last phase, the subject has to indicate where the three dots of the first phase were located in an empty 4x4 grid. The Dot Test is inspired by the Dot Memory Test presented in Sliwinski et al. 2016 [17]. More specifically, all grids were reduced in size compared to their version (5x5). We also implemented a criterion for the distractor task, namely that at least 3 F shapes are identified. If this criterion is not met, the trial is restarted. The three dots could not be on one line or in an L-shape within a 2x2 block of cells. The total score is the number of correctly indicated dots across 10 trials. This test is designed to capture visuospatial short-term memory and learning.

In the visual Backwards Digit Span (vBDS) test, a series of digits appears on the screen one by one for 1 second per digit, consistent with Hilbert et al. 2015 [18]. The subject then needs to list the digits in reversed order. Spans were randomly generated with digits between 0 and 9, using the restraints that a digit can only occur once in the span and that a chain of three or more digits could not have a fixed increment or decrement of 1 or 2, according to Woods et al. 2011 [19]. Scoring consists of calculating the sum of all correct span lengths. For example, if a subject enters two spans of length 3 and 1 of length 4 correctly, the total score is 10. This test is designed to capture working memory.

All tests were performed on a Samsung Galaxy A10 (6.2 inch screen size), with maximum sound level and screen brightness. The Symbol Test is performed in landscape position of the smartphone, whereas for the Dot Test and Backwards Digit Span, the smartphone needs to be in portrait position. In the design of ico**gnition**, careful consideration was given to the potential biasing influence of fine motor impairment in MS [13]. Motor interference was minimised by using two large buttons for the Symbol Test (contrasting digital SDMT variants where a subject has to choose from 9 smaller buttons [20]), and not putting any restrictions on the response time in the other ico**gnition** tests.

7.2.4 The validation procedure

The procedure to validate ico**gnition** is based on Benedict et al. 2012 [21], and consists of assessing four criteria addressed below.

1. Concurrent validity. Here, we assessed how well each ico**gnition** test

correlates with its paper-pencil equivalent. For the Symbol Test, this was the symbol-digit modalities test (SDMT) [10]. A sheet is presented to the subject with a key of 9 symbol-digit pairs on top and a list of symbols without a digit. In 90 seconds, the subject needs to convert as many symbols to digits as possible from the list, reading them out loud to the examiner, using the key on top. The Dot Test is based on the 10/36 spatial recall test (SPART 10/36) [22]. Here, the subject is presented a 6x6 grid with a pattern of 10 dots. This pattern is presented for 10 seconds. Subsequently, the grid is removed, and the subject is asked to replicate the pattern using 10 checkers. This cycle is repeated three times. The final score is the total number of correctly placed checkers across all trials. Finally, for the vBDS, a modified version of the WAIS-IV auditory backwards digit span test [23] was used. We will refer to this test as the "auditory Backwards Digit Span (aBDS))". Digit spans are read out loud to a subject, who is asked to repeat them in the reverse order. The original test consists of two trials of each span length, starting with a span length of 2 and incrementing with 1 each time one of the two trials is correct. As discussed in Woods et al. 2011 [24], in the original WAIS-IV design, subjects with the same score can have a different number of correct spans. To mitigate this, we used a fixed number of spans, ranging in length from 3 to 7 (table S2). The complete list was always administered for each subject. The scoring metric is the same as described earlier in the ico**gnition** Backwards Digit Span.

2. Test-retest reliability. Benedict et al. 2012 mention that this criterion should be assessed on a "small sample" of either MS patients or healthy controls [21]. We aimed to retest HC subjects 2-3 weeks after baseline testing. We used the intraclass correlation coefficient (ICC) to assess agreement between baseline and retesting. We will use the following ICC type: "two-way mixed effects, consistency, single rater/measurement" [25].

3. For each test, we compared the performance of MS subjects to those of age-, sex- and education level-matched healthy control subjects. For this, we used a Mann-Whitney U test. The analysis was repeated after correcting test performance for age, sex and education level (cfr. supplementary figure S1).

4. Lastly, the correlation of each ico**gnition** test with the Expanded Disability Status Scale (EDSS) [26], disease duration, the Beck Depression

Inventory (BDI) [27], the Fatigue Scale for Motor and Cognitive Functions (FSMC) [28], education level and age was assessed.

7.2.5 Data curation

Data were entered independently by two researchers (SD and DVL). Conflicts in data entry were resolved through mutual discussion.

7.2.6 Statistical analyses

We used an alpha level of .05 for all analyses throughout this paper. Subjects having missing data on a certain test were only excluded for that specific test. We used a Mann-Whitney U test to compare distributions without any prior assumptions with regard to the distribution of values. We used Pearson correlation (denoted "r") if both variables were normally distributed, which was assessed with a Shapiro-Wilk test. Spearman correlation (denoted "ρ") was used otherwise.

To interpret the magnitude of the test-retest reliability (ICC), we used the guidelines of Koo and Li 2016 [25]; Poor: $< .5$: Moderate: .5 – .75, Good: .75 – .9, Excellent: > 0.9.

7.3 Results

Both the MS and HC sample are described in Table 7.1.

7.3.1 Concurrent validity

Figure 7.2 shows the scatterplot of each ico**gnition** test with its paper-pencil equivalent. The Symbol Test significantly correlated with SDMT performance (HC: r = .68, P $<$ .001; MS: r = .67, P $<$ .001). There was also a significant correlation between the Dot Test and the SPART (HC: ρ = .32, P = .007; MS: ρ = .34, P = .002) and between the vBDS and its auditory equivalent (HC: ρ = .64, P $<$ .001; MS: ρ = .68, P $<$ .001).

7.3.2 Test-retest reliability

In total, 20 HC subjects were retested with an average intertest interval of 18 days (range: 14 – 23 days) after initial testing to establish test-retest reliability. One subject did not complete the dot test upon retesting Test-retest reliability was good for the symbol Test (ICC = .84, r = .85, P $<$ .001), moderate for

	People with MS	**Healthy Controls**	**p-value**
	Demographics		
N	101	82	/
Age (Mean (SD))	45.4 (10.0)	46.8 (14.7)	.59†
Sex (female:male)	63:38	54:28	.74‡
Education (Median; IQR)	15; 5	15; 4	.74†
	MS-specific		
Disease duration in years (Mean (SD))	11.7 (7.4)	/	/
Type MS (RRMS:SPMS:PPMS)	86:8:7	/	/
EDSS (Median; IQR)	3; 2	/	/
	Paper-pencil tests		
SDMT (Mean (SD))	58.5 (10.0)	58.3 (9.9)	.82†
10/36 SPART (Mean (SD))	20.6 (4.3)	20.1 (4.6)	.74†
Auditory BDS (Mean (SD))	48.4 (18.7)	46.0 (17.6)	.37†
BDI (Mean (SD)	9.3 (5.7)	5.8 (5.1)	$< .001$†
FSMC (Mean (SD))	58.7 (16.4)	39.9 (12.1)	$< .001$†
	ico**gnition** tests		
Symbol Test (Mean (SD))	24.8 (6.3)	25.4 (6.4)	.42
Dot Test (Mean (SD))	21.8 (5.1)	21.2 (4.9)	.32
vBDS (Mean (SD))	46.9 (16.9)	43.8 (16.7)	.27

Table 7.1: Group characteristics. Abbreviations: N = sample size, SD = standard deviation, IQR = Interquartile range, RRMS = Relapsing-Remitting MS, SPMS = Secondary Progressive MS, PPMS = Primary Progressive MS, EDSS = Expanded Disability Status Scale, SDMT = Symbol-Digit Modalities Test, SPART = Spatial Recall Test, BDS = Backwards Digit Span, BDI = Beck Depression Index, FSMC = Fatigue Scale for Motor and Cognitive Functions, vBDS = visual Backwards Digit Span. †Mann-Whitney U test, ‡Chi-squared test. The following variables had missing values: EDSS (7), SPART (HC: 1, MS: 1), BDI (MS: 2, HC: 1), FSMC (MS: 2, HC: 7), Dot Test (MS: 3, HC: 1), vBDS (MS: 1). An additional 3 subjects were excluded for the Dot Test since they were tested with an earlier ico**gnition** version where a larger grid size was used. We reduced the grid size after these three subjects as we deemed this test to be too difficult.

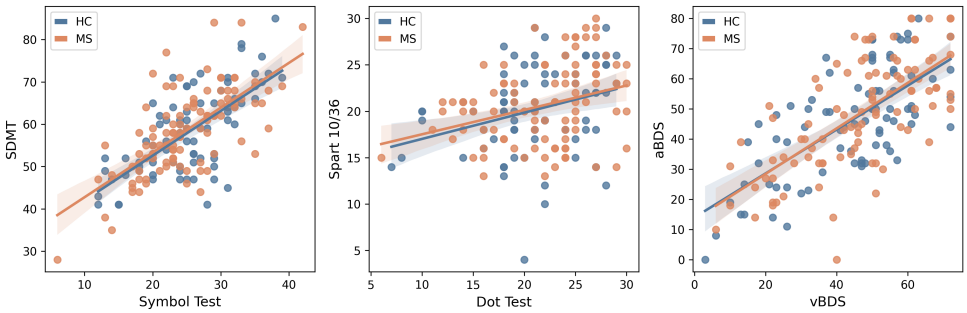

Figure 7.2: Concurrent validity. Scatterplot of each ico**gnition** test and its paper-pencil equivalent.

the Dot Test: (ICC = .73, r = .74, P < .001) and good for the vBDS (ICC = .83, ρ = .66, P = .002).

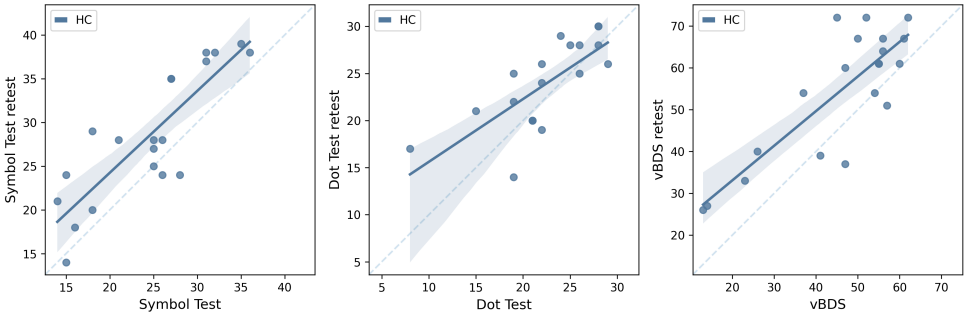

Figure 7.3: Test-retest reliability.

7.3.3 Difference MS and HC

For all tests in ico**gnition**, there was no significant difference in performance between healthy subjects and people with MS (figure 7.4). Symbol Test: U = 4431, P = .42; Dot Test: U = 3516, P = .33; vBDS: U = 3708, P = .27.

7.3.4 Correlations with clinical parameters

A correlation matrix between ico**gnition** tests on the one hand and the different clinical tests on the other can be consulted in table 7.2. In general, people with higher test scores on any of the ico**gnition** tests had less physical

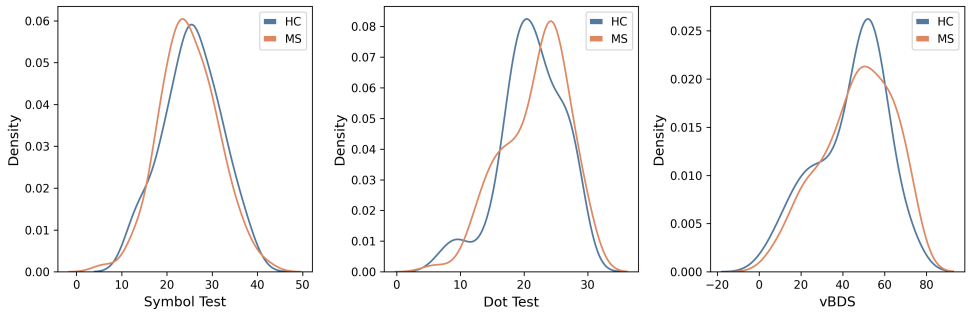

Figure 7.4: Comparison of the performance of healthy subjects and people with MS on the ico**gnition** tests.

disability (EDSS), were younger (age) and had a higher education level (number of years education). Furthermore, all but the Symbol Test correlated with disease duration, whereas fatigue (FSMC) was negatively associated with Dot Test performance. No correlation was found between any of the ico**gnition** test and depression (BDI).

	Symbol Test	Dot Test	vBDS
Age	-.52 (<.001)	-.26 (.01)	-.24 (.02)
Education level	.20 (.04)	.37 (<.001)	.22 (.03)
BDI	-.03 (.78)	-.12 (.27)	.03 (.80)
FSMC	-.13 (.20)	-.26 (.01)	-.03 (.79)
EDSS	-.34 (<.001)	-.32 (.003)	-.21 (.04)
Disease duration	-.22 (.03)	-.24 (.02)	-.29 (.003)

Table 7.2: Correlation matrix of each ico**gnition** test with several clinical variables. Each value is represented as correlation (P-value). Spearman correlation was used for all comparisons except FSMC and the Symbol Test (Pearson).

7.4 Discussion

In this paper, we present the results of the validation process of a smartphone-based screening battery for cognitive problems in persons with MS. All tests correlated with their paper-pencil equivalent (concurrent validity), clinical variables (age, education level and EDSS), and showed moderate-to-good test-retest reliability. This indicates the suitability of the battery to be used as a screening tool for routine remote follow-up of cognitive problems, which many

persons with MS will develop over time [5].

7.4.1 A cognitively preserved MS sample

Persons with MS scored equally well on all tests compared to healthy subjects. However, as can be observed in table 7.1 and supplementary figure S2, this was also the case for the paper-pencil tests. Indeed, we appear to have included an MS sample with relatively preserved cognition. We tested this in a post-hoc analysis using the analysis of variance (ANOVA) method described in Kallner et al. 2017 [29], comparing SDMT performance of our MS sample (mean = 58.5, standard deviation (SD) = 10.0, N = 101 (cfr. table 7.1)) with previous literature. López-Góngora et al. 2015 report an average SDMT performance of 54.3 (SD = 13.4, N = 237) [30], while Sousa et al. 2021 mention an average SDMT performance of 53.51 (SD = 11.76, N = 115) [31]. We found that subjects with MS included in this study performed significantly better (comparison with López-Góngora et al. 2015: F = 8.01, P = .005; comparison with Sousa et al. 2021: F = 11.1, P = .001). These results should however be assessed in light of the sample characteristics, as both studies appeared to have recruited an MS sample with lower age, higher education level, and a larger proportion of female subjects. A future study should establish the performance of ico**gnition** in a cognitively impaired MS sample matched on age, sex and education level.

7.4.2 Test-retest reliability

All ico**gnition** tests showed acceptable test-retest reliability. It can however be observed from figure 7.3 that subjects tended to score better upon retesting compared to baseline testing. We checked this post-hoc using a Wilcoxon signed-rank test between baseline and retest scores, and found a significant difference for the Symbol test (W = 15, P = .001) and the vBDS (W = 15.5, P = .001), but not for the Dot Test (W = 45.5, p = .08). We hypothesise that this is due to an initial familiarisation with the test, indicating that the trial phases of both tests might not have been sufficient. We expect this effect to gradually diminish with future testing, as all tests are characterised by randomisation.

7.4.3 Correlations with clinical tests

The ico**gnition** tests did not correlate with depression as assessed by the BDI. Although there appears to be firm evidence that depressive symptoms could impact cognitive performance [32], a recent paper by Anderson et al. 2023

reported no impact of depressive symptoms on non-executive cognitive speed and memory [33]. Although subjects with MS scored significantly higher on the BDI compared to controls in our study, they on average scored in a range characterised by no depression [34].

Lastly, only the Dot Test correlated with fatigue in MS. Subjects frequently reported this test to be difficult, which might explain why more fatigued subjects score worse on the test. An interesting future avenue would be to assess whether ico**gnition** is sensitive to cognitive fatigability, which is often reported in patients with MS [35]. Digital assessment also offers opportunities in this regard, as metadata such as reaction times can be stored such as is the case in ico**gnition**. A gradual increase in reaction time throughout cognitive tests could in turn be a proxy for cognitive fatigability.

7.4.4 Concurrent validity

The concurrent validity of the Dot Test was relatively poor with respect to the other two ico**gnition** tests. This is most likely due to the fact that the golden standard test, the SPART 10/36, does not contain a distraction phase. Furthermore, by using a True/False alternative to the SDMT, the influence of guessing might be bigger in the Symbol Test, as subjects have a 50% chance of being correct, while this is only 1/9=11% for the SDMT. Lastly, the vBDS test slightly differed from the aBDS test, as the vBDS included an element of randomisation and the additional restrictions that repetition of a digit was not allowed, nor a fixed increment of 1 or 2 in a succession of 3 digits.

7.4.5 The benefits of regular digital follow-up

Proper and regular follow-up of cognitive function is important to capture fluctuations in cognitive state, for example due to a disease exacerbation or relapse. Although Giedraitiene et al. 2018 show that a short cognitive screening tool like BICAMS [11] can pick up these cognitive fluctuations [36], the cognitive dip is likely missed in reality as cognitive assessments are usually done once or twice a year during clinical follow-up visits. ico**gnition** allows for more frequent cognitive assessments. Furthermore, testing can be performed wherever and whenever a patient feels ready for it, reducing biasing effects such as fatigue [37]. Digitalisation moreover allows to extract more information from a cognitive test, such as cognitive fatigue by tracking response time in the Symbol Test. By implementing ico**gnition** in a health care platform with symptom logging and patient-reported outcome measures (PROMS) [38],

patients take an active role in their disease management; they can provide the individual information that complements the professional knowledge of the treating physician [39].

7.4.6 Limitations and future work

There are some limitations to this study. In the initial version of the Dot Test, the final 5 trials included remembering the position of the three dots on a 5x5 grid. After testing 3 subjects with MS, we however decided to consistently use a 4x4 grid given the difficulty of the task, and excluded these subjects from the dot test analyses. Including these 3 subjects in a post-hoc analysis did, however, yield similar results. We furthermore tested consistently on the same Android smartphone to reduce biasing effects such as different screen size and weight. However, as ico**gnition** is designed using Flutter, it can also be deployed on other operating systems than Android. An external validation study is planned to show robustness of the app on different devices in a home setting.

7.5 Conclusion

This study presents a newly developed smartphone app, ico**gnition**. It was found to be a valid and reliable tool for remote screening of cognitive functioning in persons with MS. This allows for regular follow-up of cognitive performance to more quickly respond to deterioration.

7.6 Availability of data and code

The anonymised source data on which this paper relies, as well as the code used for statistical analysis and the creation of figures, will be made available in our lab's GitHub repository: https://github.com/AIMS-VUB/smartphone_tests.

7.7 Supplementary material

7.7.1 Z-normalisation

Procedure

To normalise the ico**gnition** test scores for age, sex (male: 1, female: 2) and education level (years of education), for each ico**gnition** test, we fitted a linear regression equation with the test performance as dependent variable and age, sex and education level as independent variables. We then predicted the expected test score of HC and MS subjects given the healthy control regression equation, and subtracted the predicted value from the true test score, yielding the prediction error (ϵ_{pred}).

$$test\_score_{pred} = w_0 + w_1 * sex + w_2 * education\_level + w_3 * age \tag{7.1}$$

$$\epsilon_{pred} = test\_score_{norm} - test\_score_{pred} \tag{7.2}$$

For each subject, the z score is the prediction error normalised with respect to the standard deviation of the prediction error distribution of the healthy control subjects.

$$z = \frac{\epsilon_{pred}}{std(\epsilon_{pred,HC})} \tag{7.3}$$

Necessary values

The necessary information to perform the procedure above is included below.

	Symbol Test	Dot Test	Backwards Digit Span
w0	29.8987	28.8806	18.4677
w1	0.2170	-1.5233	0.6054
w2	0.5795	0.1559	2.4597
w3	-0.2849	-0.1610	-0.2506
std($\epsilon_{pred,HC}$)	4.4741	4.1522	15.2010

Table S7.1: Values for the normalisation procedure per ico**gnition** test

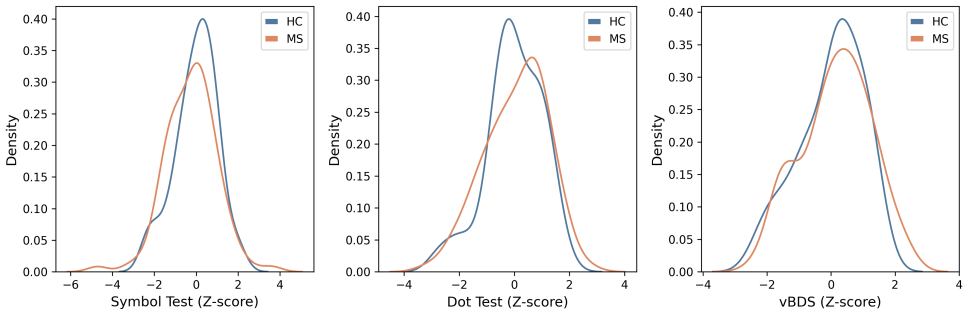

Figure S7.1: Performance on normalised test scores of persons with MS and healthy controls.

7.7.2 Criterion validity on normalised test scores

Figure S7.1 displays the performance of each ico**gnition** test after correcting test performance for expected performance from the healthy control dataset. People with MS and healthy controls (HC) had comparable test scores (Symbol Test (Z): U = 4367, p = 0.192; Dot Test (Z): U = 3780, p = 0.683; vBDS (Z): U = 3623, p = 0.386).

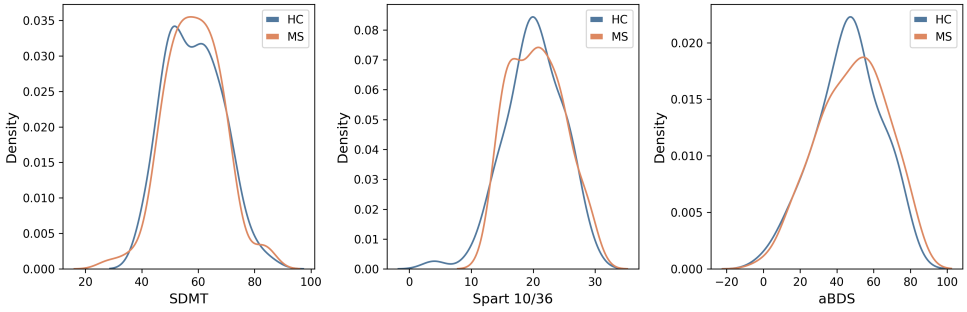

Figure S7.2: Performance on paper-pencil tests of persons with MS and healthy controls.

7.7.3 Test performance MS versus HC: paper-pencil tests

None of the paper-pencil tests were significantly different between people with MS and HC subjects (SDMT: U = 3794.5, p = 0.715; SPART 10/36: U = 3792, p = 0.709; auditory Digit Span Backwards (aBDS): U = 3568.5, p = 0.304). The criterion validity (MS versus HC) for all paper-pencil test equivalents of

all ico**gnition** tests can be consulted in figure S7.2.

7.7.4 Digit spans

The digit spans used in the auditory Backwards Digit Span are listed in table S7.2.

Span length	Digit span
3	927
3	843
4	3164
4	5360
4	7918
4	4895
5	28452
5	26019
5	30475
5	84857
6	213604
6	918536
6	639271
6	728938
7	2395480
7	5837686

Table S7.2: Digit spans used in the auditory backwards digit span

References

[1] Denissen, S., Van Laethem, D., Baijot, J., Costers, L., Descamps, A., Van Remoortel, A., Van Merhaegen-Wieleman, A., D'hooghe, M.B., D'Haeseleer, M., Smeets, D. et al. icognition: a smartphone-based cognitive screening battery. *medRxiv*, pages 2023–07, 2023.

[2] Yeroushalmi, S., Maloni, H., Costello, K. and Wallin, M.T. Telemedicine and multiple sclerosis: A comprehensive literature review. *J Telemed Telecare*, 26(7-8):400–413, August 2020.

[3] Khan, F., Amatya, B., Kesselring, J. and Galea, M. Telerehabilitation for persons with multiple sclerosis. *Cochrane Database Syst Rev*, 2015(4):CD010508, April 2015.

[4] Charvet, L.E., Yang, J., Shaw, M.T., Sherman, K., Haider, L., Xu, J. and Krupp, L.B. Cognitive function in multiple sclerosis improves with telerehabilitation: Results from a randomized controlled trial. *PLoS One*, 12(5):e0177177, 2017.

[5] Ruano, L., Portaccio, E., Goretti, B., Niccolai, C., Severo, M., Patti, F., Cilia, S., Gallo, P., Grossi, P., Ghezzi, A. et al. Age and disability drive cognitive impairment in multiple sclerosis across disease subtypes. *Mult Scler*, 23(9):1258–1267, August 2017.

[6] DeLuca, J., Chiaravalloti, N.D. and Sandroff, B.M. Treatment and management of cognitive dysfunction in patients with multiple sclerosis. *Nature Reviews Neurology*, 16(6):319–332, 2020.

[7] Macías Islas, M.A. and Ciampi, E. Assessment and Impact of Cognitive Impairment in Multiple Sclerosis: An Overview. *Biomedicines*, 7(1):22, March 2019. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Foong, Y.C., Bridge, F., Merlo, D., Gresle, M., Zhu, C., Buzzard, K., Butzkueven, H. and van der Walt, A. Smartphone monitoring of cognition in people with multiple sclerosis: A systematic review. *Multiple Sclerosis and Related Disorders*, page 104674, 2023.

[9] Beier, M., Alschuler, K., Amtmann, D., Hughes, A., Madathil, R. and Ehde, D. iCAMS: Assessing the Reliability of a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) Tablet Application. *International Journal of MS Care*, 22(2):67–74, March 2020.

[10] Benedict, R.H., DeLuca, J., Phillips, G., LaRocca, N., Hudson, L.D. and Rudick, R. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. *Mult Scler*, 23(5):721–733, April 2017.

[11] Langdon, D., Amato, M., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hämäläinen, P., Hartung, H.P., Krupp, L., Penner, I. et al. Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *Mult Scler*, 18(6):891–898, June 2012.

[12] Pereira, D.R., Costa, P. and Cerqueira, J.J. Repeated Assessment and Practice Effects of the Written Symbol Digit Modalities Test Using a Short Inter-Test Interval. *Arch Clin Neuropsychol*, 30(5):424–434, August 2015.

[13] Lamers, I., Kerkhofs, L., Raats, J., Kos, D., Van Wijmeersch, B. and Feys, P. Perceived and actual arm performance in multiple sclerosis: relationship with clinical tests according to hand dominance. *Mult Scler*, 19(10):1341–1348, September 2013.

[14] Van Hecke, W., Costers, L., Descamps, A., Ribbens, A., Nagels, G., Smeets, D. and Sima, D.M. A Novel Digital Care Management Platform to Monitor Clinical and Subclinical Disease Activity in Multiple Sclerosis. *Brain Sciences*, 11(9):1171, September 2021. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[15] Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S. et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2):162–173, February 2018. Publisher: Elsevier.

[16] Rypma, B., Berger, J.S., Prabhakaran, V., Martin Bly, B., Kimberg, D.Y., Biswal, B.B. and D'Esposito, M. Neural correlates of cognitive efficiency. *NeuroImage*, 33(3):969–979, November 2006.

[17] Sliwinski, M.J., Mogle, J.A., Hyun, J., Munoz, E., Smyth, J.M. and Lipton, R.B. Reliability and Validity of Ambulatory Cognitive Assessments. *Assessment*, 25(1):14–30, January 2018.

[18] Hilbert, S., Nakagawa, T.T., Puci, P., Zech, A. and Bühner, M. The digit span backwards task: Verbal and visual cognitive strategies in working memory assessment. *European Journal of Psychological Assessment*, 31:174–180, 2015. Place: Germany Publisher: Hogrefe Publishing.

[19] Woods, D.L., Herron, T.J., Yund, E.W., Hink, R.F., Kishiyama, M.M. and Reed, B. Computerized analysis of error patterns in digit span recall. *J Clin Exp Neuropsychol*, 33(7):721–734, August 2011.

[20] Pham, L., Harris, T., Varosanec, M., Morgan, V., Kosa, P. and Bielekova, B. Smartphone-based symbol-digit modalities test reliably captures brain damage in multiple sclerosis. *npj Digit. Med.*, 4(1):1–13, February 2021. Number: 1 Publisher: Nature Publishing Group.

[21] Benedict, R.H., Amato, M.P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hamalainen, P., Hartung, H., Krupp, L., Penner, I. et al. Brief International Cognitive Assessment for MS (BICAMS): international standards for validation. *BMC Neurology*, 12(1):55, July 2012.

[22] Gerstenecker, A., Martin, R., Marson, D.C., Bashir, K. and Triebel, K.L. Introducing Demographic-Corrections for the 10/36 Spatial Recall Test. *Int J Geriatr Psychiatry*, 31(4):406–411, April 2016.

[23] Raiford, S.E., Coalson, D.L., Saklofske, D.H. and Weiss, L.G. CHAPTER 2 - Practical Issues in WAIS-IV Administration and Scoring. In Weiss, L.G., Saklofske, D.H., Coalson, D.L. and Raiford, S.E., editors, *WAIS-IV Clinical Use and Interpretation*, Practical Resources for the Mental Health Professional, pages 25–59. Academic Press, San Diego, January 2010.

[24] Woods, D.L., Kishiyama, M.M., Yund, E.W., Herron, T.J., Edwards, B., Poliva, O., Hink, R.F. and Reed, B. Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, 33(1):101–111, January 2011. Publisher: Routledge _eprint: https://doi.org/10.1080/13803395.2010.493149.

[25] Koo, T.K. and Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*, 15(2):155–163, June 2016.

[26] Kurtzke, J.F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1452, November 1983.

[27] Beck, A.T., Ward, C.H., Mendelson, M., Mock, J. and Erbaugh, J. An inventory for measuring depression. *Arch Gen Psychiatry*, 4:561–571, June 1961.

[28] Penner, I.K., Raselli, C., Stöcklin, M., Opwis, K., Kappos, L. and Calabrese, P. The Fatigue Scale for Motor and Cognitive Functions (FSMC): validation of a new instrument to assess multiple sclerosis-related fatigue. *Mult Scler*, 15(12):1509–1517, December 2009.

[29] Kallner, A. Resolution of Students t-tests, ANOVA and analysis of variance components from intermediary data. *Biochem Med (Zagreb)*, 27(2):253–258, June 2017.

[30] López-Góngora, M., Querol, L. and Escartín, A. A one-year follow-up study of the Symbol Digit Modalities Test (SDMT) and the Paced Auditory Serial Addition Test (PASAT) in relapsing-remitting multiple sclerosis: an appraisal of comparative longitudinal sensitivity. *BMC Neurology*, 15(1):40, March 2015.

[31] Sousa, C., Rigueiro-Neves, M., Passos, A.M., Ferreira, A. and Sá, M.J. Assessment of cognitive functions in patients with multiple sclerosis applying the normative values of the Rao's brief repeatable battery in the Portuguese population. *BMC Neurol*, 21(1):1–10, December 2021. Number: 1 Publisher: BioMed Central.

[32] Margoni, M., Preziosa, P., Rocca, M.A. and Filippi, M. Depressive symptoms, anxiety and cognitive impairment: emerging evidence in multiple sclerosis. *Translational Psychiatry*, 13(1):264, 2023.

[33] Anderson, J.R., Fitzgerald, K.C., Murrough, J.W., Katz Sand, I.B., Sorets, T.R., Krieger, S.C., Riley, C.S., Fabian, M.T. and Sumowski, J.F. Depression symptoms and cognition in multiple sclerosis: Longitudinal evidence of a specific link to executive control. *Multiple Sclerosis Journal*, 29(13):1632–1645, 2023.

[34] Dozois, D.J., Dobson, K.S. and Ahnberg, J.L. A psychometric evaluation of the Beck Depression Inventory–II. *Psychological assessment*, 10(2):83, 1998.

[35] Morrow, S.A., Rosehart, H. and Johnson, A.M. Diagnosis and quantification of cognitive fatigue in multiple sclerosis. *Cognitive and Behavioral Neurology*, 28(1):27–32, 2015.

[36] Giedraitiene, N., Kaubrys, G. and Kizlaitiene, R. Cognition during and after multiple sclerosis relapse as assessed with the brief international cognitive assessment for multiple sclerosis. *Scientific reports*, 8(1):8169, 2018.

[37] Andreasen, A.K., Iversen, P., Marstrand, L., Siersma, V., Siebner, H.R. and Sellebjerg, F. Structural and cognitive correlates of fatigue in progressive multiple sclerosis. *Neurological research*, 41(2):168–176, 2019.

[38] Montalban, X., Graves, J., Midaglia, L., Mulero, P., Julian, L., Baker, M., Schadrack, J., Gossens, C., Ganzetti, M., Scotland, A. et al. A smartphone sensor-based digital outcome assessment of multiple sclerosis. *Mult Scler*, 28(4):654–664, April 2022.

[39] Holman, H. and Lorig, K. Patients as partners in managing chronic disease: partnership is a prerequisite for effective and efficient health care, 2000.

Chapter 8

Brain age as a surrogate marker for cognitive performance in MS

Stijn Denissen[1,2], Denis Alexander Engemann[3,4], Alexander De Cock[1], Lars Costers[1,2], Johan Baijot[1], Jorne Laton[1,5], Iris-Katharina Penner[6,7], Matthias Grothe[8], Michael Kirsch[9], Marie Beatrice D'hooghe[10,11], Miguel D'Haeseleer[10], Dominique Dive[12], Johan De Mey[13], Jeroen Van Schependom[1,14], Diana Maria Sima[1,2], Guy Nagels[1,2,15]

1 AIMS Lab, Center for Neurosciences, UZ Brussel, VUB, Brussels, Belgium **2** Icometrix, Leuven, Belgium **3** Université Paris-Saclay, Inria, CEA, Palaiseau, France **4** Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany **5** Nuffield Department of Clinical Neurosciences, University of Oxford, Headley Way, Oxford, UK **6** Cogito Center for Applied Neurocognition and Neuropsychological Research, Düsseldorf, Germany **7** Department of Neurology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany **8** Department of Neurology, University Medicine Greifswald, Greifswald, Germany **9** Institute for Diagnostic Radiology and Neuroradiology, University Medicine of Greifswald, Greifswald, Germany **10** National Multiple Sclerosis Center Melsbroek, Melsbroek, Belgium **11** Center for Neurosciences, VUB, Brussels, Belgium **12** Department of Neurology, University Hospital of Liege, Esneux, Belgium **13** Department of Radiology, UZ Brussel, Brussels, Belgium **14** Department of Electronics and Informatics (ETRO), VUB, Brussels, Belgium **15** St Edmund Hall, University of Oxford, Queen's Lane, Oxford, UK

This chapter is based on a paper in the *European Journal of Neurology* [1]

Abstract

Background and purpose: Data from neuroimaging techniques allow us to estimate a brain's age. Brain age is easily interpretable as 'how old the brain looks' and could there-fore be an attractive communication tool for brain health in clinical practice. This study aimed to investigate its clinical utility by investigating the relationship between brain age and cognitive performance in multiple sclerosis (MS).

Methods: A linear regression model was trained to predict age from brain magnetic resonance imaging volumetric features and sex in a healthy control dataset (HC_train, n = 1673). This model was used to predict brain age in two test sets: HC_test (n = 50) and MS_test (n = 201). Brain-predicted age difference (BPAD) was calculated as BPAD = brain age minus chronological age. Cognitive performance was assessed by the Symbol Digit Modalities Test (SDMT).

Results: Brain age was significantly related to SDMT scores in the MS_test dataset (r = -0.46, $p < 0.001$) and contributed uniquely to variance in SDMT beyond chronological age, reflected by a significant correlation between BPAD and SDMT (r = -0.24, $p < 0.001$) and a significant weight (-0.25, p = 0.002) in a multivariate regression equation with age.

Conclusions: Brain age is a candidate biomarker for cognitive dysfunction in MS and an easy to grasp metric for brain health.

Keywords
multiple sclerosis | cognition | biomarkers | machine learning | magnetic resonance imaging | brain age

8.1 Introduction

About half of the people with multiple sclerosis (MS) experience cognitive impairment [2], aggravating the impact of MS on their daily life and that of their caregivers [3]. Susceptibility to cognitive impairment should be assessed holistically, as it depends on clinical factors such as age [4], disability [4], premorbid intelligence [4] and disease duration [5] but also on findings in other domains such as brain imaging [6]. The most prominent of these cognitive difficulties is a slowing of information processing abilities [7], which literature suggests to be the key driver for deficits in other cognitive domains in MS [8]. Timely identification of cognitive difficulties is imperative, as it allows for early treatment planning (in particular cognitive rehabilitation appears to be effective, although other methods exist [9]) to both prevent and address patient-specific difficulties associated with cognitive impairment. These include falling, reduced quality of life and employment issues [8], but their impact extends to the mental health of caregivers [10]. Early detection requires regular and consistent follow-up in standard clinical care. Currently, neuropsychological testing remains the gold standard to detect cognitive problems [11], the most popular test in clinical practice being the Symbol Digit Modalities Test (SDMT) [12]. Although the SDMT is a brief screening test [13], it is prone to practice effects [14].

An objective biomarker to diagnose cognitive deficits might circumvent the aforementioned problem. Currently, predominantly structural brain characteristics, extracted from brain imaging techniques such as magnetic resonance imaging (MRI), were found to be related to cognitive performance [15]. Yet more information can be extracted from an MRI than meets the eye. By using large datasets of brain images of healthy individuals, a machine learning model can be trained to estimate the age of a given brain. For a new brain image, the model will output the best guess of the age of that per-son's brain, that is, the 'brain age', which can look older or younger than the actual, chronological age of that person. The elegance of brain age lies in its interpretable nature; it is easily graspable how old a brain appears to be. In several brain disorders [16], including MS [16, 17, 18], brains typically look older than those of their healthy peers.

On top of being an interpretable metric, recent evidence indicates that brain age could explain clinical symptoms in MS, reflected by statistically significant, albeit weak, correlations. More specifically, increased brain age is associated with physical disability as quantified by the Expanded Disability

Status Scale (EDSS, r = 0.23) [16] and the nine-hole peg test (r = 0.36) [17]. Beyond physical dis-ability, recent findings by Kaufmann et al. [16] in dementia suggest that increased brain age could explain cognitive disability as well, namely by being associated with lower scores on the Mini Mental State Examination (r = -0.30), independent of chronological age.

In summary, brain age is an interpretable imaging- derived metric that is sensitive to MS-related pathology. However, it is currently unknown how brain ageing is related to MS-specific cognitive dysfunction. So far, efforts have mostly uncovered anatomical correlates by directly linking brain volumetry to cognitive performance. Yet, people might be more easily capable of imagining 'how old a brain looks' compared to 'how voluminous a brain is', posing an opportunity for a new communication tool in clinical practice that avoids medical jargon, which answers the desire of patients to be informed in plain language [19]. In this study, brain age is explored as a tool for studying cognitive dysfunction in MS on an international, multi-centre dataset.

8.2 Methods

8.2.1 Data description

Data are described by subdividing them into HC_train, HC_test and MS_test, used to train and test the brain age decoding model, as shown in figure 8.1. HC_train was constructed from a large sample of 1673 healthy control (HC) subjects from online publicly available repositories (only subjects of 18 years or older were included, consistent with the training dataset of Cole et al. [20]). Refer to table S8.1 and figure S8.1 for a more detailed description. For the test datasets, two centres contributed retrospective data to this study. First, in Brussels, both HC (n = 50) and MS (n = 97) subjects were assessed as part of a study [21, 22] on understanding the neural origins of cognitive disturbances in MS. The MS subjects of this study were recruited at the National Multiple Sclerosis Center of Melsbroek. Second, the Universitätsmedizin Greifswald contributed data from 104 MS subjects to this study. Altogether, this resulted in 50 subjects for the HC_test and 201 subjects for the MS_test. For all data, T1- weighted MRI, sex and age at image acquisition were available. For the test datasets, fluid attenuated inversion recovery (FLAIR) MRI and results from the SDMT [12] were available. Finally, MS_test data also contained EDSS, disease duration and type of MS. A summary of all data is available in table 8.1.

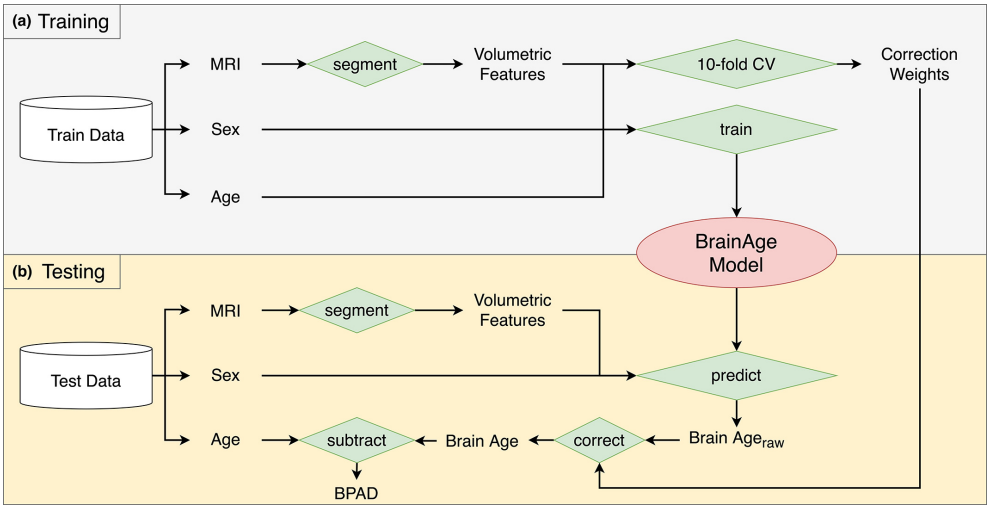

Figure 8.1: Brain age pipeline. The pipeline is subdivided into (a) a training phase and (b) a testing phase, where 'Train Data' refers to the HC_train data and 'Test Data' represents either the HC_test dataset or the MS_test dataset. A silo-like shape represents a dataset, whereas green diamonds represent some kind of operation, specified by the text. Other text represents either variables or images

The symbol-digit modalities test (SDMT). We chose the SDMT, explained in chapter 2, as a proxy for cognitive functioning. It is a brief test, designed to measure information processing speed, that is attractive for its psychometric properties [23], quick administration [13] and its capability of predicting scores on other cognitive tests [24]. Moreover, compared to other cognitive tests, the SDMT is the first to flag cognitive deterioration [24]. As a recent study by Sandry et al. [25] found that SDMT performance is not solely determined by one single cognitive process, that is, information processing speed, its results were interpreted as measuring global cognitive performance.

Brain MRI. Brain age research mostly uses T1-weighted brain MR images to model upon. Plausibly, this is mainly driven by data availability; open source repositories with healthy ageing data mostly contain T1w brain images, allowing the creation of large imaging databases, enabling to robustly train a brain age model. In MS, using T1w brain images allows expressing damage caused by neurodegeneration, the main disease process in MS besides neuroinflammation, in terms of ageing. This study therefore also targeted the neurodegenerative component of MS.

Dataset	**HC_train**	**HC_test**		**MS_test**	
Source	Public	Brussels	Brussels	Greifswald	Total
			Data description		
N	1673	50	97	104	201
Age					
- Mean ± SD	41.9 ± 19.5	48.0 ± 11.9	48.1 ± 9.6	43.1 ± 12.0	45.5 ± 11.2
- Range (min–max)[b]	18–94	26–68	26–70	20–69	20–70
Gender (M:F)	673:1000	19:31	29:68	35:69	64:137
EDSS (median; IQR)	-	-	3.0; 2.0	1.5; 2.0	2.5; 2.5
Disease duration (years)	–	–	15.7 ± 8.4	8.4 ± 6.2	11.9 ± 8.2
MS subtype	-	-	CIS: 2	CIS: 0	CIS: 2
	-	-	RRMS:82	RRMS: 100	RRMS: 182
	-	-	SPMS:6	SPMS: 1	SPMS: 7
	-	-	PPMS:7	PPMS: 3	PPMS: 10
SDMT (mean ± SD)	-	53.8 ± 9.6	48.0 ± 11.4	51.2 ± 15.0	49.6 ± 13.5
			Scanner description		
Field strength (T)	1.5 and 3[a]	3	3	3	3
Sequences	T1	T1 + FLAIR	T1 + FLAIR	T1 + FLAIR	T1 + FLAIR
Scanner	Various[a]	Philips Ingenia: 36 Philips Achieva: 14	Philips Ingenia: 68 Philips Achieva: 29	Siemens Verio	Philips Ingenia: 68 Philips Achieva: 29 Siemens Verio: 104

Table 8.1: Data characteristics. Abbreviations: CIS, clinically isolated syndrome; EDSS, Expanded Disability Status Scale; F, female; FLAIR, fluid attenuated inversion recovery; IQR, interquartile range; M, male; MS, multiple sclerosis; PPMS, primary progressive MS; RRMS, relapsing–remitting MS; SDMT, Symbol Digit Modalities Test; SPMS, secondary progressive MS. [a] Refer to table S8.1 for more details. [b] Values displayed as integer ages (rounded down).

8.2.2 Ethics

All participants of the BRUMEG study, Brussels, provided their written informed consent prior to MRI assessment. The study pro-tocol (B.U.N. 143201423263) was approved by the ethical com-mittee of the UZ Brussel (Commissie Medische Ethiek [O.G. 016], Reflectiegroep Biomedische Ethiek) on 25 February 2015. For the patients from Greifswald, Germany, the study was approved by the ethics committee of the Medical Faculty of the University of Greifswald (BB159/18), and all participants gave their written in-formed consent. HC_train data consist of publicly available data originating from other projects, listed in table S8.1. Ethical approval was received by each project separately.

8.2.3 Magnetic resonance imaging preprocessing and brain age pipeline

Several preparatory steps were involved in the construction of our brain age model. They are summarised below.

1. *Brain MRI segmentation.* From the HC_train dataset, 3D T1-weighted MRI, sex and chronological age were extracted. Next, the T1-weighted MR images were evaluated by the FDA cleared icobrain software (version 4.4) of icometrix NV (Leuven, Belgium). This is an end-to- end automatic software that segments and subsequently quantifies distinct regions of the brain, which was originally published in Jain et al. [26]. The pipeline relies on T1-weighted MR images and, when available, also uses FLAIR images to segment white matter lesions in the brain. These lesions are filled in the T1 image with white matter intensities, and the T1 image is subsequently segmented by fitting a probabilistic model for grey matter, white matter and cerebrospinal fluid image intensities. Ultimately, the pipeline generates volumes of the segmented regions, yielding a set of features, normalised for head size, that describe the brain's morphology. These can be subdivided into general volumes (grey matter, white matter, lateral ventricles), lobe-specific cortical grey matter (frontal, temporal, parietal, occipital) and subcortical volumes (hippocampus [left and right] and thalamus [left and right]). Together with the subjects' sex, this forms the total set of features used for the brain age pipeline. HC_test and MS_test were also segmented with ico**brain**, yielding the aforementioned set of features. The sole difference with HC_train is the additional availability of FLAIR images to perform lesion filling in the T1 image.

2. *Linear regression.* Along with chronological age, that is, the target variable to be predicted, the total set of features (z-normalised) served as input for training a supervised machine learning model. Ordinary least squares regression was used, since it is amongst the most interpretable machine learning algorithms; the predicted brain age is simply the weighted sum of all features with their respective weight:

$$brain\ age = w_0 + w_1 feature_1 + w_2 feature_2 + ... + w_n feature_n$$

with w0 being the intercept or 'bias'. The training phase consists of finding the weights that minimise the error between the prediction, that is, brain age, and a subject's true age. To ensure the comparability of weights, each feature was normalised as follows:

$$feature_{normalized} = \frac{feature - mean(feature_{HC\_train})}{std(feature_{HC\_train})}$$

3. *Brain age correction.* According to Le et al. [27], age-predicting models are prone to 'regression towards the mean', a phenomenon that results in overestimation of the brain age of younger subjects and underestimation of the brain age of older subjects. With the example of a linear brain age model, Smith et al. 2019 report that brain age is not orthogonal to, and therefore dependent on, age [28]. The dependency increases when model performance drops [28]. There appears to be firm consensus throughout the literature that such bias should be corrected for, and several methods exist to do so [29]. The method described by Cole et al. [18] was used. First, BrainAge_raw on HC_train was estimated by adopting 10-fold cross-validation (CV). Second, a linear regression equation was fitted between the obtained raw brain ages and the respective chronological ages:

$$BrainAge_{raw} = \beta_0 + \beta_1 ChronologicalAge + \epsilon$$

Here, β_0 and β_1 represent the intercept and slope of the regression line, respectively, whereas the error term ϵ represents the residuals between data points and the regression line. β_0 and β_1 serve to correct each brain age predicted by our model using

$$BrainAge = (BrainAge_{raw} - \beta_0)/\beta_1$$

They were first used to correct the raw brain age estimates of HC_train.

In summary, after segmentation of the MR images, 10-fold CV on HC_train was first performed to obtain the correction weights, which additionally allows the mean absolute error (MAE) to be calculated on the HC_train dataset, providing an intuition in model performance on HC_train. Secondly, all available training data were used to train a final brain age model that was used for further analyses. Altogether, this is referred to as the training phase, which is represented visually in figure 8.1a. Next, whether the learned weights generalise to other datasets as well was investigated (figure 8.1b). Raw brain age on the HC_test and MS_test datasets was first calculated by calculating the weighted sum of the features (i.e., sex and brain volumes from ico**brain**) with their respective weights, which were then corrected by using the correction formula of preparatory step 3. In the remainder of this paper, 'brain age' consistently refers to the corrected brain age.

Finally, the latter variable was additionally used to calculate the brain-predicted age difference (BPAD) [18]. It quantifies brain age overestimation by subtracting chronological age from brain age.

8.2.4 Statistical analyses

Statistical analyses and visualisations were performed in Python and R. Significance level alpha was set to 0.05 for all reported test results. Pearson correlation was used. Raincloud plots were generated with use of the Ptit-Prince package [30].

Non-parametric tests were used to compare distributions without making any assumption about the distribution the samples were drawn from. First, a Mann–Whitney U test was used to compare brain age, BPAD and chronological age distributions between MS_test and HC_test. To compare the BPAD of both test sets with 0, a one-sample Wilcoxon signed rank test was used.

Second, the error of predicting age from brain images was calculated with MAE between true and predicted age for the healthy control datasets. Next, the Pearson correlation was used to establish the association between brain age and SDMT in the MS_test data. To assess whether it contains unique information beyond chronological age, two approaches were used. First, brain age and chronological age were considered together in a multivariate linear regression equation:

Feature	Weight	Standard Error	t	p
Intercept	41.8867	0.216	193.509	<0.001
Grey matter	-7.5859	1.021	-7.427	<0.001
White matter	-0.8746	0.269	-3.252	0.001
Lateral ventricles	2.2432	0.359	6.244	<0.001
Cortical grey matter—frontal lobe	-4.0462	0.628	-6.438	<0.001
Cortical grey matter—occipital lobe	0.4299	0.314	1.370	0.171
Cortical grey matter—temporal lobe	-1.0626	0.403	-2.638	0.008
Cortical grey matter—parietal lobe	-1.0037	0.424	-2.366	0.018
Hippocampus—left	0.2280	0.304	0.750	0.453
Hippocampus—right	1.4228	0.311	4.573	<0.001
Thalamus—left	-2.1227	0.526	-4.036	<0.001
Thalamus—right	-2.5859	0.511	-5.057	<0.001
Sex	-3.4668	0.229	-15.136	<0.001

Table 8.2: The final brain age model's characteristics.

$$SDMT = \beta_0 + \beta_1 BrainAge + \beta_2 ChronologicalAge + \epsilon$$

Second, the relationship between BPAD and SDMT in the MS_test data was assessed using a Pearson correlation.

Predicting brain age from brain volumetry and sex using linear regression essentially represents a linear transformation that reduces the dimensionality of the feature space from 12 (brain volumetry + sex) to 1 (brain age). Another type of linear transformation that is commonly used to compress a set of variables is principal component analysis (PCA). PCA essentially 'reorganises' the variables to a set of principal components that are uncorrelated and explains variance in a dataset in decreasing order; the first principal component (PC_1) explains the majority of this variance. To construct PC_1, a PCA was first fitted on the feature space of the HC_train dataset. Next, this was used to transform the feature space of the MS_test dataset by projecting along PC_1. The relationship between brain age and PC_1 in the MS_test data was assessed with a Pearson correlation.

8.3 Results

8.3.1 The brain age pipeline

Linear regression between brain age, obtained with 10-fold CV on HC_train (cf. figure 8.1a), and chronological age yielded the following correction weights:

Dataset	**HC_train**	**HC_test**		**MS_test**	
Source	Public	Brussels	Brussels	Greifswald	Total
N	1673	50	97	104	201
Brain age (mean ± SD)	41.9 ± 21.9†	46.1 ± 16.8	61.8 ± 16.6	62.6 ± 22.9	62.2 ± 20.1
BPAD (mean ± SD)	0 ± 10.0†	-1.9 ± 9.7	13.7 ± 14.7	19.5 ± 16.0	16.7 ± 15.6

Table 8.3: Outputs from the brain age pipeline: corrected brain age and BPAD for the HC_train, HC_test and MS_test datasets. Note: The dagger (†) indicates that these values were obtained by means of 10-fold cross-validation. Abbreviation: BPAD, brain-predicted age difference.

$\beta_0 = 8.60$, $\beta_1 = 0.79$. In the same dataset, the MAE between corrected brain age and chronological age was 7.91 years. The result of the brain age correction is added to the supplementary material as figure S8.2.

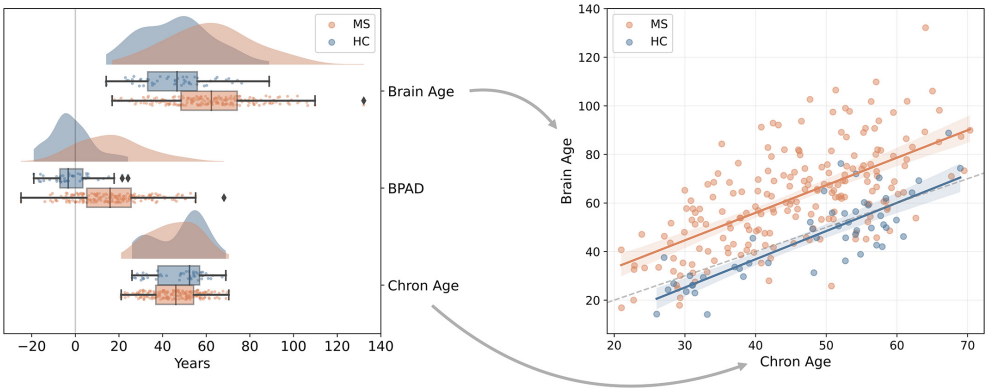

Figure 8.2: Group comparison between HC_test (blue) and MS_test (orange) for brain age, BPAD and chronological age. Left: The raincloud plots show the distribution of brain age, BPAD and chronological age for MS_test and HC_test. A reference line at x= 0 is included as visual aid. Right: The scatterplot shows the relationship between brain age and chronological age for MS_test ($r = 0.63$, $p < 0.001$) and HC_test ($r = 0.82$, $p < 0.001$). The dotted line is added as reference, namely where brain age = chronological age

.

In a next step, a final brain age model was fitted by using all available HC_train data for training (cf. figure 8.1a). The model's feature weights can be consulted in table 8.2. To test its quality, it was applied to the HC_test set and MS_test set, represented in figure 8.1b. The corrected brain age and BPAD values are summarised in table 8.3 and visually displayed for HC_test and MS_test in the raincloud plots of figure 8.2.

1. Evaluation on HC_test data. The model predicted brain age with an MAE of 7.85 years. BPAD was not significantly different from zero (T = 449, p = 0.069), indicating that, on average, brain age is similar to chronological age.

2. Evaluation on MS_test data. BPAD was significantly greater than zero (T = 1121, p < 0.001). It is additionally noted that both brain age (U = 2708, p < 0.001) and BPAD (U = 1506, p < 0.001) were significantly higher in the MS_test data compared to HC_test. Chronological age was comparable between the two groups (U = 5721, p = 0.130).

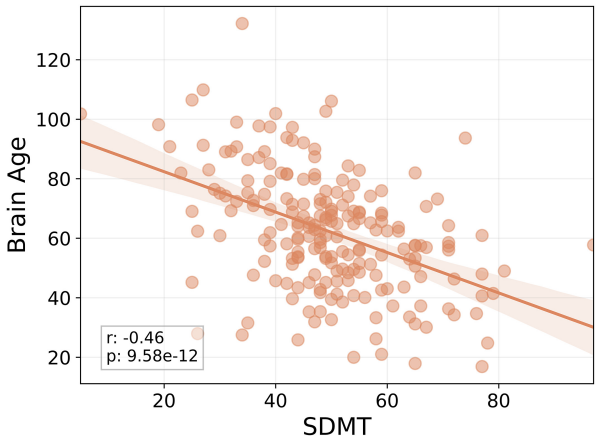

Figure 8.3: Scatterplot between brain age and SDMT in the MS_test dataset. The textbox describes the Pearson r statistic, along with the p value

8.3.2 The relation between brain age and cognitive performance

Brain age was significantly correlated with SDMT (figure 8.3, r = -0.46, p < 0.001) and explained 20.85% of the variance in SDMT (R^2). Moreover, brain age explained unique variance in SDMT beyond chronological age, which is reflected by a significant correlation between BPAD and SDMT (figure 8.4, left, r = -0.24, p < 0.001) and the significant weight (β_1 = -0.25, p = 0.002) assigned to brain age when considering it in the multivariate regression equation (figure 8.4, right). Chronological age also contributed significantly to the model (β_2 = -0.32, p < 0.001).

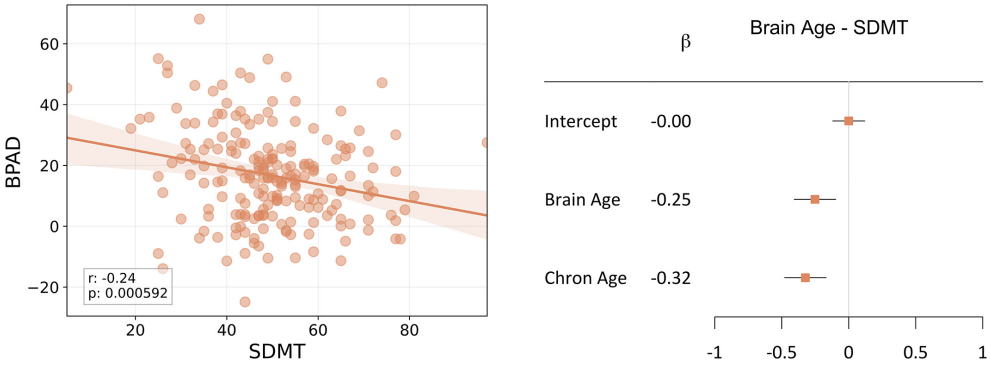

Figure 8.4: The relationship between brain age and SDMT, independent of chronological age. Left: Scatterplot between BPAD and SDMT in the MS_test dataset. The textbox describes the Pearson r statistic, along with the p value. Right: Forest plot visualising the significance of the weights (β_n) in the linear regression equation $SDMT = \beta_0 + \beta_1 BrainAge + \beta_2 ChronologicalAge + \epsilon$ in the MS_test dataset (note that variables were normalised with respect to mean and standard deviation before input in the regression equation). The maximum likelihood estimates of the weights (β_n) are represented by the orange squares, along with a 95% confidence interval (horizontal bar). If the latter does not include 0, the contribution of that feature to the model is considered significant. Brain age and chronological age contributed significantly (p = 0.002 and p < 0.001 respectively)

8.3.3 The relation between brain age and brain volumetry

Figure 8.5 displays the relationship between PC_1 and brain age, revealing a strong linear relationship (r = 0.93, p < 0.001).

Table S8.3 shows the correlation of each brain volumetric feature with brain age. Whole brain volume, normalized for head size, was also included (r = -0.92, p < 0.001).

8.3.4 The relation between brain age and other clinical variables

Brain age was significantly correlated with both EDSS (r = 0.37, p < 0.001) and disease duration (r = 0.32, p < 0.001). BPAD was significantly correlated with EDSS (r = 0.17, p = 0.018) but not with disease duration (r = 0.04, p = 0.586). Figure S8.6 shows the effect of EDSS and disease duration, as well as age, on the relationship between brain age and SDMT.

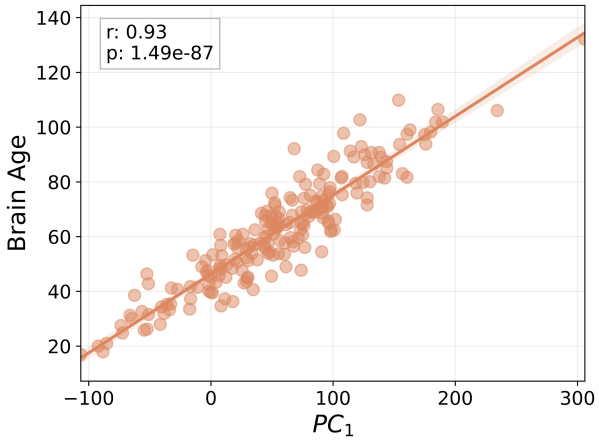

Figure 8.5: Scatterplot between the first principal component (PC_1) and brain age in the MS_test dataset. The text box describes the Pearson r statistic, along with the p value

8.4 Discussion

This study aimed to investigate the potential of brain age, an intuitive metric of brain health, as a biomarker for cognitive dysfunction in MS. Our results suggest that brain age could be a promising candidate; it is significantly related to cognitive performance, independent of chronological age. Moreover, it was shown that brain age explained the majority of variance in brain volumetry by establishing a strong relationship with the PC_1 of our total set of features; both linear transformations appear to yield a similar metric of brain health.

8.4.1 Brain age and brain-predicted age difference (BPAD)

In the past few years, most brain-age-related research in MS was dedicated to establishing clinical correlates of the BPAD [16, 17, 18]. Although BPAD might be regarded as a simplification of brain age, valuable information is in fact lost by subtracting two variables. Our results support this statement in two ways. First, BPAD showed a weaker correlation with cognitive performance than brain age. Second, brain age and chronological age both contributed unique information in explaining cognitive performance in MS. In terms of clinical significance, BPAD could be valuable for monitoring patients over time, for example to assess treatment effect. Since it takes into account the age at MRI assessment, a reduction of BPAD at follow-up might indicate decreased disease activity, for example as a result of a certain treatment.

8.4.2 Brain age compared to existing biomarkers

Although brain age is a fair choice for decoding cognitive performance in MS, it is noted that its performance was comparable to whole brain volume, as shown by Golan et al. [31], reporting a correlation of r = 0.46 between whole brain volume and global cognitive functioning. This is consistent with findings of other studies with large sample sizes, reporting a higher prevalence of cognitive impairment in subjects with lower brain parenchymal fraction [32] and a correlation of r = 0.50 between whole brain fraction and processing speed [33]. Nonetheless, brain age has an important advantage in contrast to any biological correlate of cognition in MS: it is easy to grasp as 'how old a brain looks', which facilitates communication with persons with MS. Constructing an uncomplicated message with minimal jargon contributes to optimally transferring medical information to patients, in turn optimising patient care [19]. The flip-side of the same coin, however, might be that because patients can better imagine this metric of brain health it could be traumatic if not carefully communicated. The suitability of the metric as a communication tool will probably differ from patient to patient and should be carefully considered when aiming for a more personalised approach to medicine. Future research on patients' attitude towards brain age might shed new lights on patients' acceptance of brain age.

8.4.3 User trust

However, an important hurdle in the path of brain age models to clinical practice is nicely illustrated by a statement in Ribeiro et al.: 'if the users do not trust a model or a prediction, they will not use it' [34]. This issue should be addressed, in particular as efforts emerge to include brain age models in routine MRI examination [35]. First, to maximise trust of the MS clinician in our model, simple linear regression was used. The MS clinician has been familiarised with this method by decades of research adopting it for various purposes, for example studying the relation between MRI and cognitive performance in MS [36, 37]. The advantage of a linear model is that the impact of a change in a feature (in our case brain volumes and sex) is directly observable in the brain age: the change multiplied by the weight equals the number of years the brain will be estimated younger or older (the sign of the weight indicates whether the brain age will be estimated younger or older, whereas the magnitude of the weight indicates the number of years). Linear models are therefore both interpretable (obvious causal relationship between input and output) [38] and explainable (good understanding of the model's

internal mechanisms) [38]. This contrasts with other studies on brain age in MS, mostly adopting models that are common in machine learning but not in clinical research, such as Gaussian processes regression [18] and extreme gradient boosting [17]. Secondly, trust in the prediction of our model, that is, predicted brain age, was enhanced by showing that brain age explains the majority of the variance in MRI-derived volumetric features and sex, used to train our brain age model. This observation is logical in the light of what is known about the ageing brain, shrinking as people get older [39].

8.4.4 Interpretation of regression weights in the brain age model

As mentioned in the previous section, the weights in table 8.2 represent the relative contribution of each feature to the estimation of brain age. All volumetric variables were normalised with respect to head size, and each variable was normalised with respect to the mean and standard deviation of the respective feature on the HC_train data set. The sex variable was coded as 0 (Female) and 1 (Male), which was converted by the normalisation procedure to 1.219 and -0.820 respectively (rounded to 3 decimals). Since sex was assigned a negative value, being male yields a reduction in brain age compared to being female. The reason for this could be that males have a larger total brain volume compared to women [40]. The volumetric variables were predominantly assigned a negative weight, with the largest weight being assigned to grey matter volume. Together with a positive weight for lateral ventricles volume (indicating loss of brain tissue), this aligns with the expectation that preservation of brain tissue results in a lower brain age estimation. Grey matter volume was likely assigned the highest weight in the linear equation since it decreases approximately linearly over time after the age of 6 [41, 42]. This can also be observed in figure 8.6, where the change of each volumetric feature with age on the HC_train data set was plotted.

Cortical grey matter of the occipital lobe and hippocampal volume were assigned a positive weight, albeit small. Although surprising for cortical grey matter, judging from figure 8.6, hippocampal volume only starts decreasing around middle age. The positive weight for occipital lobe cortical grey matter could be explained by not using a regularisation term while multicollinearity was present between the volumetric features, which might have confused the model in assigning weights. This is discussed in more detail in the limitations section.

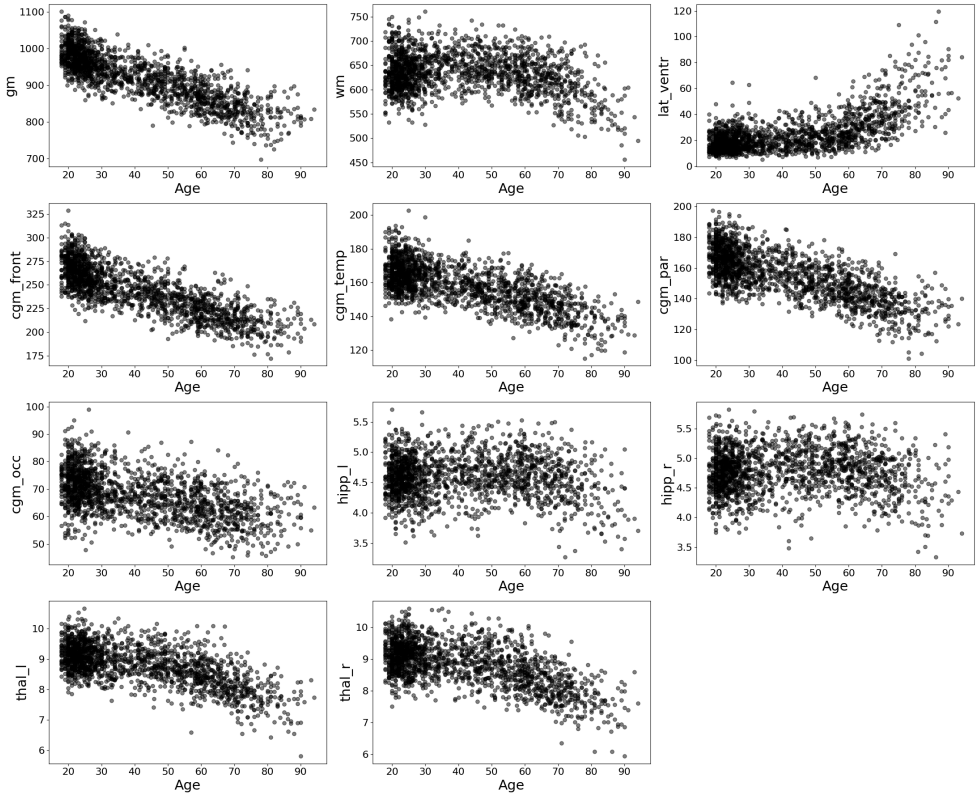

Figure 8.6: Change of each volumetric feature with age on the HC_train data set. For more information on the abbreviations used as y labels, we refer to table 8.2.

8.4.5 Model performance and clinical implications

Our model achieved an MAE of 7.91 years on the HC_train dataset (10-fold CV) and 7.85 years on an independent HC test set. This is relatively large compared to previously published models that adopted a more complex methodology compared to ours. For example, Cole et al. [18] achieved an MAE of 5.02 years on their training sample. However, as brain age is foremost a surrogate marker for clinical variables of interest, models should be assessed in terms of their clinical utility, rather than focusing solely on their age decoding capacity. The brain age model of Cole et al. [18] was therefore applied to our data, relying on Gaussian processes regression and being publicly available. A detailed description of the methodology and results of this post hoc analysis is available in the supplementary material. The model of Cole et al. [18] achieved an MAE of 5.52 on our HC_test sample, and a Pearson correlation

between brain age and chronological age of 0.89 ($p < 0.001$) and 0.78 ($p < 0.001$) respectively for the HC_test and MS_test data set. An F test was used to compare the variance of the BPAD distributions of the Cole model and our model, indicating that the Cole model was significantly more accurate in decoding age compared to our model ($F = 2.23$, $p = 0.006$). Interestingly, however, the correlations with SDMT of brain age (figure S8.4, $r = -0.50$, $p < 0.001$) and BPAD (figure S8.5, $r = -0.21$, $p = 0.003$) obtained with the model of Cole et al. [18] were similar to those of brain age ($r = -0.46$, $p < 0.001$) and BPAD ($r = -0.24$, $p < 0.001$) resulting from our model. Differences in correlation coefficients were not statistically significant (brain age, $z = -1.43$, $p = 0.153$; BPAD, $z = 0.69$, $p = 0.492$).

Hence, although the models differ with respect to performance in age decoding, the models are comparable in terms of clinical significance. As one should always strive to make models as simple as possible [43], our model is deemed to be more suitable for clinical practice. In terms of benchmarking, future research could however investigate the performance of an even simpler model, i.e. one single feature such as grey matter volume, in predicting brain age.

8.4.6 Unique variance of brain age beyond chronological age

This manuscript used two ways to correct brain age for chronological age, assessing the unique variance in SDMT explained by brain age beyond chronological age. First, we correlated BPAD with SDMT, and second, included brain age and chronological age together in a linear regression equation with SDMT. The latter approach might be criticised for not removing an inherent dependency of both brain age and SDMT on chronological age. Moreover, Smith et al. 2019 highlight that brain age estimations are biased by a non-linear dependency on chronological age [28]. We therefore performed a post-hoc analysis where we first fit two linear regression equations. In both equations, both chronological age and chronological age squared were included as independent variables, whereas the dependent variable was either corrected brain age or SDMT. We then extracted the residuals of both equations, and assessed those together in a new linear regression equation. The latter resulted in a significant relationship between both residuals (weight: -0.2841, standard error: 0.09094, t: -3.125, p: .002). We moreover repeated this analysis for the corrected brain age of the model of Cole et al. 2019 (cfr. previous section), also yielding a significant relationship between residuals (weight: -0.1707, standard error: 0.05282, t: -3.233, p: .001).

8.4.7 The consistent use of Pearson correlation

In this manuscript, we consistently used Pearson correlation to assess the relationship between two variables. We however note that some of these correlations were calculated on variables that were not normally distributed (assessed with a Shapiro-Wilk test [44]), which included chronological age for the HC_test sample, and brain age (Cole model), BPAD (Cole model), chronological age, EDSS, disease duration, whole brain volume, white matter volume, lateral ventricles volume, right hippocampus volume and left and right thalamus volume for the MS_test sample. Using Spearman correlation for comparisons including these variables yielded similar results, except for the correlation between EDSS and BPAD in the MS_test sample, which is non-significant when using Spearman correlation (0.10, p = 0.175).

8.4.8 Limitations

Our results imply that brain age has the potential to explain cognitive status in people with MS; brain age explained 20.85% (R^2) of the variance in SDMT. Yet, the cross-sectional nature of the data limited us to investigate the potential of brain age to predict future cognitive decline. Furthermore, previous literature highlights the importance of paying careful attention when using data from different scanners for brain age research [45]. One way to address this issue is to maximise the variety of scanners used in the training set, which might prevent brain age models from becoming highly dependent on a specific type of scanner. This was the case for our HC_train dataset (cf. table S8.1 for data sources), in which all three scanner types that were used in the test datasets were also represented (Philips Achieva, Philips Ingenia and Siemens Verio). Nonetheless, preprocessing of brain images already partly counteracts heterogeneity across scanners, as the ico**brain** software used to segment the MR images shows limited inter-scanner variability [46, 47]. Therefore, this bias has been deemed to have been properly addressed.

In this study, we modelled the ageing brain with simple linear regression. The rationale for using linear regression is its interpretable nature, and as reported in Bethlehem et al. 2022, several regional brain volumes approximate linear trajectories throughout ageing [41]. Judging from the same paper, however, linear regression is likely an oversimplification as variables predominantly follow non-linear trajectories. This means that the assumption of linear regression that relationships between dependent and independent variables are linear [48] has not been met. Two key assumptions for linear regression [48]

have been addressed below:

- **Linear relationship between the dependent variable and each independent variable.** This condition was not met, as volumetric features follow non-linear trajectories, even after the age of 18 [41]. Figure 8.6 shows the change of each volumetric feature with age on the HC_train data set. The non-linear relationships might explain the aforementioned non-linear dependency of our brain age model on chronological age.
- **The independent variables are linearly independent of each other.** As can be observed in the heat map of figure 8.7, significant multicollinearity is present between the independent variables of the model. We however did not include a regularisation term. Regularisation such as L1 (LASSO) or L2 (Ridge) bounds the model's weights, therefore preventing that a feature receives an excessive weight, which might indicate overfitting. In an earlier version of this manuscript, we in fact included an L2 regularisation [49], but removed this after removing several summary features such as total cortical grey matter volume, as regional cortical grey matter volume was already present in the feature representation. Judging from table 8.2, no feature was assigned an excessively large weight after removing these summary features and the regularisation term. Nonetheless, these feature weights should be interpreted with caution, as the features are not independent of each other. An increase in the "frontal cortical grey matter" variable will for example also cause an increase in the "grey matter" variable, thereby obfuscating the pure contribution of cortical grey matter to the model.

As can be observed in figure 8.1, the age distribution of the HC_test data is a bimodal distribution. It is important to keep this observation in mind when interpreting the performance of the model on unseen test data, especially in light of the non-linear dependency of brain age on chronological age as reported before. The model achieved an MAE of 7.91 years, and had an overall tendency to underestimate age (mean BPAD = -1.9 years). Future studies might consider testing their models on a sample with a uniform distribution that represents all ages equally.

Lastly, T1-weighted brain images were used in light of data availability. Other imaging modalities, such as T2-weighted brain images, might however better reflect other ageing processes, such as the deposition of iron, which has also been reported in MS [50]. Moreover, neuroinflammatory damage caused

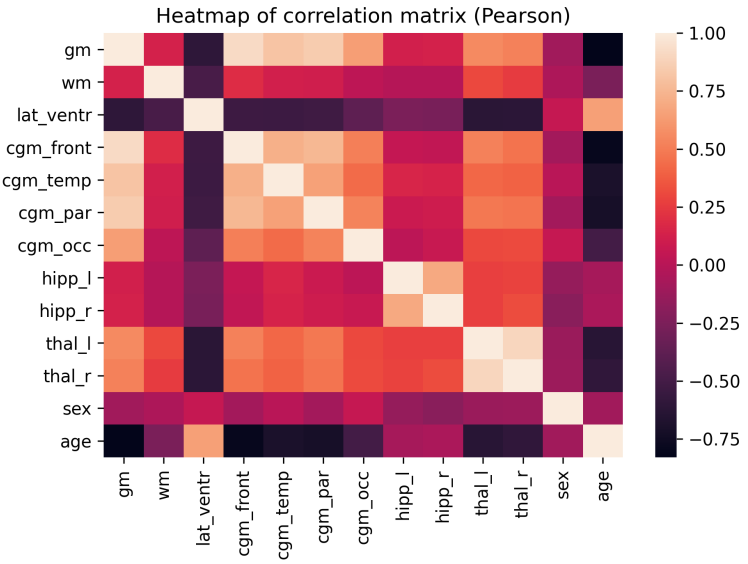

Figure 8.7: Heat map of a correlation matrix (Pearson) of the features of the brain age model (cfr. table 8.2).

by MS is better visible on FLAIR images. The latter however might have little implications for the concept of brain age, as neuroinflammatory damage caused by MS is not expected to occur in healthy ageing; brain age is most likely mainly sensitive to the neurodegenerative aspect of MS.

8.4.9 Conclusive statement

In summary, the methodology of a linear brain age model can be interpreted by clinicians and its prediction by patients. Together with its potential to explain cognitive performance, predicted brain age could be a valuable clinical tool to analyse and communicate results from brain imaging data in MS.

8.5 Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

8.6 Supplementary material

8.6.1 HC_train data characteristics

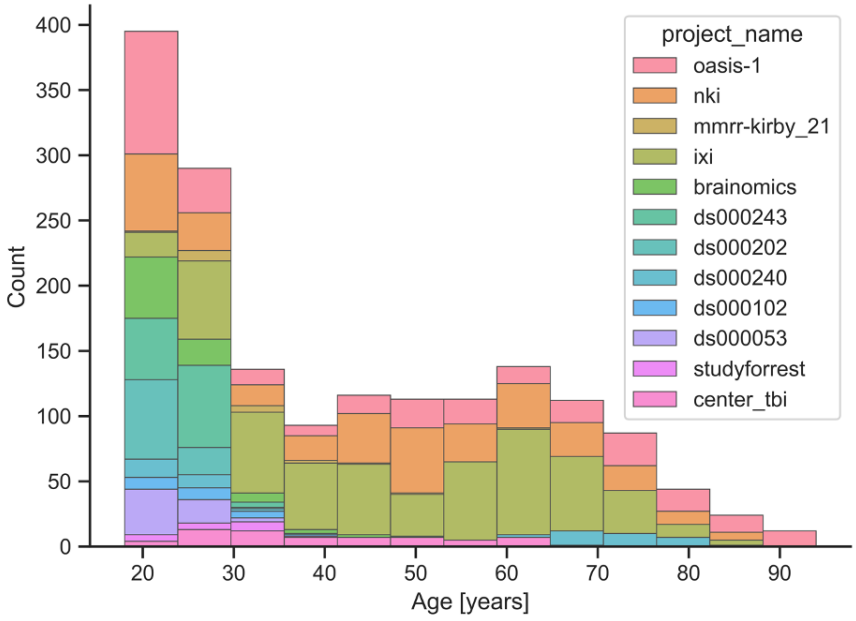

Figure S8.1: Histogram of the chronological ages per data source in the HC_train dataset

Source	N	Sex (M:F)	Scanner (Field strength (Tesla))	Additional info	Reference
studyforrest	18	10:8	Philips Achieva (3T)	http://studyforrest.org/access.html	[51]
oasis-1	300	113:187	Siemens Vision (1.5T)	http://www.oasis-brains.org/	[52]
nki	335	112:223	Siemens TrioTim (3T)	http://fcon_1000.projects.nitrc.org/indi/pro/nki.html	[53]
mmrr-kirby_21	19	10:9	Philips Achieva (3T)	http://www.nitrc.org/frs/?group_id=313	[54]
ixi	523	235:288	Philips Intera (3T) Philips Gyroscan Intera (1.5T) GE (1.5T)	http://brain-development.org/ixi-dataset/	[55]
ds000243	114	57:57	Siemens TrioTim (3T)	https://openneuro.org/datasets/ds000243/versions/00001 https://www.openfmri.org/dataset/ds000243/	[56]
ds000240	58	25:33	Siemens TrioTim (3T)	https://openneuro.org/datasets/ds000240/versions/00002	[57]
ds000202	83	0:83	Philips Achieva (3T)	https://openneuro.org/datasets/ds000202/versions/00001	[58]
ds000102	24	14:10	Siemens Allegra (3T)	https://openneuro.org/datasets/ds000102/versions/00001	[59]
ds000053	56	27:29	Siemens Skyra (3T)	https://openneuro.org/datasets/ds000053/versions/00001	[60]
center_tbi	63	33:30	Various (3T)	https://www.center-tbi.eu/	[61]
brainomics	80	37:43	Siemens Triotim (3T) Bruker (3T)	https://hal-cea.archives-ouvertes.fr/cea-01213448	[62]

Table S8.1: Characteristics of each data source in the HC_train dataset.

8.6.2 Brain age correction

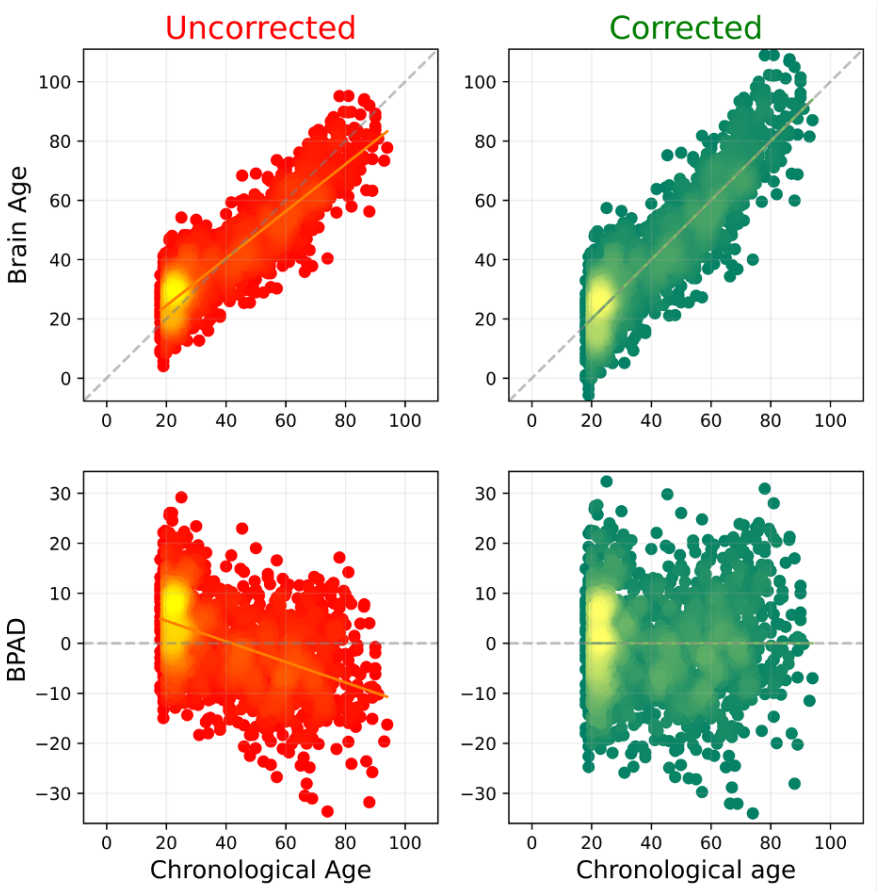

Figure S8.2: Brain age and BPAD as calculated with 10-fold cross-validation on the HC_train data. Uncorrected (raw) brain age and BPAD are indicated in red (regression line is indicated in orange), whereas corrected values are shown in green (regression line is indicated in light green). Both colour maps indicate higher density when yellower. Dotted lines were added to each plot as visual reference. For the upper plots, this is the line where Brain Age = Chronological Age, whereas for the lower plots, this is the line where BPAD = 0. Before correction, the regression lines do not coincide with the dotted lines, resulting in an overestimation of brain age in young individuals and an underestimation in older subjects. This is no longer the case after correction.

8.6.3 Comparison with the brain age model of Cole et al. 2020 [18]

This section describes the results of applying the model of Cole et al. 2020 [18] to the HC_test and MS_test dataset. Each T1-weighted MR image was processed by the model, which is publicly available at https://github.com/james-cole/brainageR. As this brain age was not yet corrected (we will refer to the uncorrected brain age as the "raw brain age"), we used the correction procedure as described in Cole et al. 2020 [18] (cfr. equation below).

$$BrainAge = \frac{(BrainAge_{raw} - 3.33)}{0.91}$$

Brain age and BPAD distributions

Brain age and BPAD resulting from the Cole model are listed in table S8.2.

	Brain Age			BPAD		
	Cole model	Our model	Test result	Cole model	Our model	Test result
HC_test	45.91 ± 14.38	46.08 ± 16.76	616 (0.836)	-2.05 ± 6.50	-1.88 ± 9.71	616 (0.836)
MS_test	53.18 ± 14.47	62.23 ± 20.06	2312 (<.001)	7.65 ± 9.13	16.71 ± 15.59	2312 (<.001)

Table S8.2: Brain age and BPAD resulting from applying the Cole brain age model to our test datasets. Values in the "model" columns are represented as Mean ± Standard Deviation, whereas the test result is represented as Wilcoxon Test Statistic (p-value). Significant test results ($p < .05$) are presented in bold. HC_test: $n = 50$, MS_test: $n = 201$.

Model performance comparison

HC_test

The distributions of BPAD on HC_test of the Cole model, as well as our model, are shown in figure S8.3. The Cole model achieved a mean absolute error (MAE) of 5.52 years, whereas our model achieved an MAE of 7.85 years. To establish whether the Cole model significantly outperformed our model in terms of age decoding, we tested the difference in variance between both distributions with an F-test, after establishing that the BPAD distribution of our model ($W = 0.96$, $p = 0.082$) and the model of Cole ($W = 0.97$, $p = 0.304$) were normally distributed with a Shapiro-Wilk test. The F-test revealed a significant difference in BPAD variance between our model and the model of Cole ($F = 2.23$, $p = 0.006$). Hence, the brain age estimations of the model of Cole were significantly more accurate compared to the brain age estimations

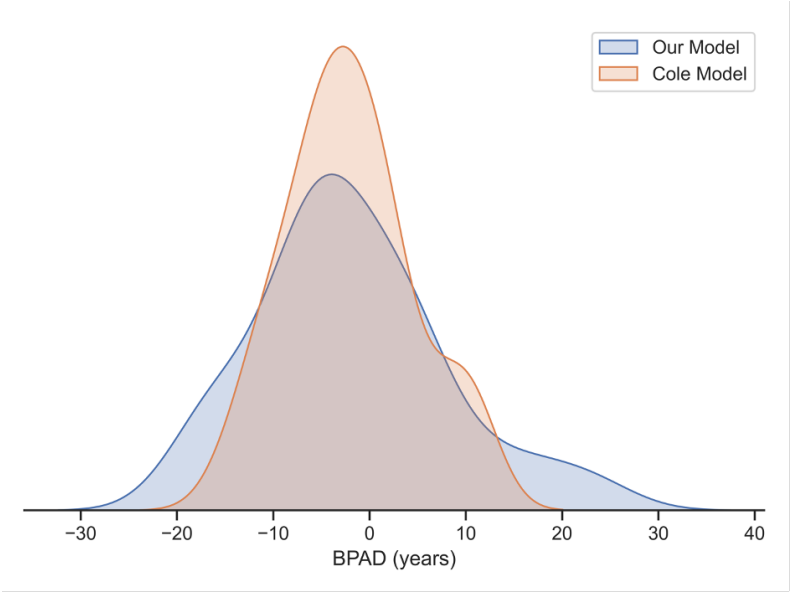

Figure S8.3: BPAD distributions of both models on HC_test.

of our model.

MS_test
On the MS_test set, both brain age and BPAD were significantly lower for the Cole model compared to our model. This observation is most likely attributable to the difference in input of both models; our model uses fluid-attenuated inversion recovery (FLAIR) images to estimate lesion volume, which are subsequently added to the white matter volume, whereas they are most likely segmented as grey matter in T1 images.

The relation of brain age and BPAD with cognitive performance

Cognitive performance was significantly correlated to both brain age ($r = -0.50$, $p < .001$, figure S8.4) and BPAD ($r = -0.21$, $p = 0.003$, figure S8.5) in the MS_test dataset. To compare the correlation coefficients with the correlation coefficients that resulted from our model, we used the "cocor" package in R by Diedenhofen and Musch 2015 [63]. For both brain age and BPAD, we used the "cocor.dep.groups.overlap()" function since for both models, correlations were established on the same dataset (MS_test) and using a common variable (SDMT). In this function, we set the "test" argument to "pearson1898", which

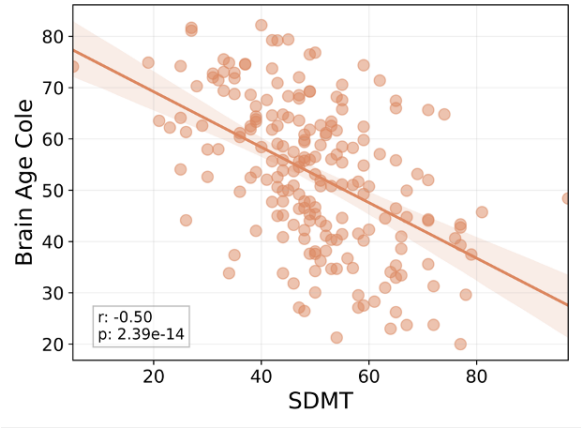

Figure S8.4: Scatterplot between brain age resulting from the Cole model and SDMT on the MS_test dataset.

uses the method from Pearson and Filon 1898 [64]. Correlation coefficients with SDMT were comparable between both models for brain age (z = -1.43, p = 0.153) and BPAD (z = 0.69, p = 0.492).

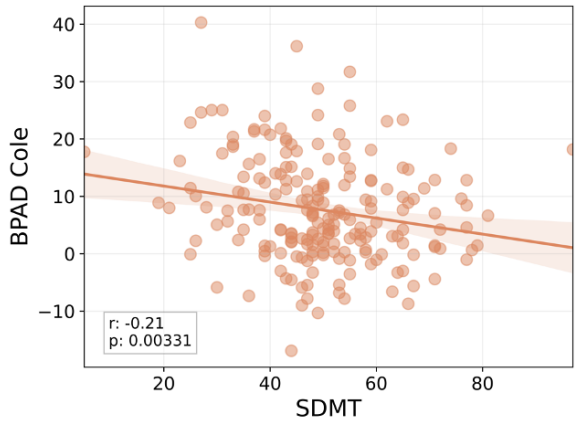

Figure S8.5: Scatterplot between BPAD resulting from the Cole model and SDMT on the MS_test dataset.

8.6.4 The relation between brain volumetric features and brain age

table S8.3 shows the correlations between the brain volumetric features that served as input features for our model (and additionally whole brain volume) and brain age.

	brain age
whole brain	-0.92 (<.001)
grey matter	-0.90 (<.001)
white matter	-0.53 (<.001)
lateral ventricles	0.73 (<.001)
cortical grey matter – frontal lobe	-0.79 (<.001)
cortical grey matter – occipital lobe	-0.54 (<.001)
cortical grey matter – temporal lobe	-0.69 (<.001)
cortical grey matter – parietal lobe	-0.79 (<.001)
hippocampus - left	-0.29 (<.001)
hippocampus - right	-0.29 (<.001)
thalamus - left	-0.73 (<.001)
thalamus - right	-0.68 (<.001)

Table S8.3: Correlations of brain volumetric features with brain age. Values are expressed as: Pearson r (p value).

8.6.5 The effect of age, EDSS and disease duration on the relationship between brain age and SDMT

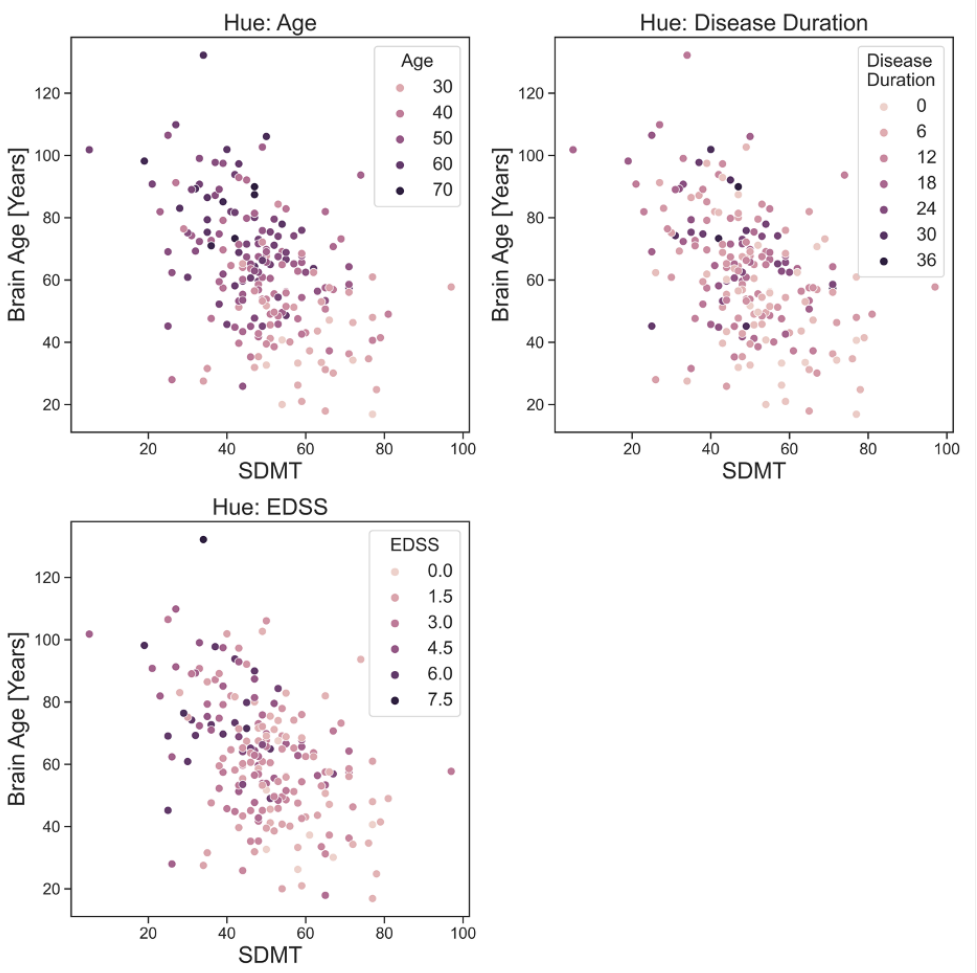

Figure S8.6: Scatterplots between SDMT and brain age, hued on age (upper left), disease duration (upper right) and EDSS (lower left).

References

[1] Denissen, S., Engemann, D.A., De Cock, A., Costers, L., Baijot, J., Laton, J., Penner, I.K., Grothe, M., Kirsch, M., D'hooghe, M.B. et al. Brain age as a surrogate marker for cognitive performance in multiple sclerosis. *European journal of neurology*, 29(10):3039–3049, 2022.

[2] Rao, S.M., Leo, G.J., Bernardin, L. and Unverzagt, F. Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. *Neurology*, 41(5):685–91, may 1991.

[3] Islas, M.Á.M. and Ciampi, E. Assessment and impact of cognitive impairment in multiple sclerosis: An overview, mar 2019.

[4] Amato, M.P., Prestipino, E., Bellinvia, A., Niccolai, C., Razzolini, L., Pastò, L., Fratangelo, R., Tudisco, L., Fonderico, M., Mattiolo, P.L. et al. Cognitive impairment in multiple sclerosis: An exploratory analysis of environmental and lifestyle risk factors. *PLOS ONE*, 14(10):e0222929, oct 2019.

[5] Brochet, B. and Ruet, A. Cognitive Impairment in Multiple Sclerosis With Regards to Disease Duration and Clinical Phenotypes. *Frontiers in Neurology*, 10:261, mar 2019.

[6] Filippi, M., Rocca, M.A., Benedict, R.H., Deluca, J., Geurts, J.J., Rombouts, S.A., Ron, M. and Comi, G. The contribution of MRI in assessing cognitive impairment in multiple sclerosis. *Neurology*, 75(23):2121, dec 2010.

[7] Chiaravalloti, N.D. and DeLuca, J. Cognitive impairment in multiple sclerosis. *The Lancet Neurology*, 7(12):1139–1151, dec 2008.

[8] Costa, S.L., Genova, H.M., Deluca, J. and Chiaravalloti, N.D. Information processing speed in multiple sclerosis: Past, present, and future, may 2017.

[9] DeLuca, J., Chiaravalloti, N.D. and Sandroff, B.M. Treatment and management of cognitive dysfunction in patients with multiple sclerosis. *Nature Reviews Neurology 2020 16:6*, 16(6):319–332, may 2020.

[10] Labiano-Fontcuberta, A., Mitchell, A.J., Moreno-García, S. and Benito-León, J. Anxiety and depressive symptoms in caregivers of multiple sclerosis patients: The role of information processing speed impairment. *Journal of the Neurological Sciences*, 349(1-2):220–225, feb 2015.

[11] Kalb, R., Beier, M., Benedict, R.H., Charvet, L., Costello, K., Feinstein, A., Gingold, J., Goverover, Y., Halper, J., Harris, C. et al. Recommendations for cognitive screening and management in multiple sclerosis care. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 24(13):1665–1680, nov 2018.

[12] Smith, A. Symbol digit modalities test: Manual. Los Angeles: Western Psychological Services. Technical report, 1982.

[13] Langdon, D.W., Amato, M.P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hämäläinen, P., Hartung, H.P., Krupp, L., Penner, I.K. et al. Recommendations for a brief international cognitive assessment for multiple sclerosis (BICAMS), 2012.

[14] Portaccio, E., Goretti, B., Zipoli, V., Iudice, A., Pina, D.D., Malentacchi, G.M., Sabatini, S., Annunziata, P., Falcini, M., Mazzoni, M. et al. Reliability, practice effects, and change indices for Raos brief repeatable battery. *Multiple Sclerosis*, 16(5):611–617, may 2010.

[15] Van Schependom, J. and Nagels, G. Targeting cognitive impairment in multiple sclerosis-the road toward an imaging-based biomarker. *Frontiers in Neuroscience*, 11(JUN):380, jun 2017.

[16] Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A. et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, 22(10):1617–1623, oct 2019.

[17] Høgestøl, E.A., Kaufmann, T., Nygaard, G.O., Beyer, M.K., Sowa, P., Nordvik, J.E., Kolskår, K., Richard, G., Andreassen, O.A., Harbo, H.F. et al. Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis. *Frontiers in Neurology*, 10(APR), 2019.

[18] Cole PhD, J.H., Raffel MD, J., Friede PhD, T., Eshaghi MD, PhD, A., Brownlee PhD, FRACP, W.J., Chard MD, PhD, D., De Stefano MD, PhD, N., Enzinger MD, C., Pirpamer MSc, L., Filippi MD, FEAN, M. et al. Longitudinal Assessment of Multiple Sclerosis with the Brain-Age Paradigm. *Annals of Neurology*, 88(1):93–105, jul 2020.

[19] King, A. and Hoppe, R.B. "Best Practice" for Patient-Centered Communication: A Narrative Review. *Journal of Graduate Medical Education*, 5(3):385–393, sep 2013.

[20] Cole, J.H., Underwood, J., Caan, M.W., De Francesco, D., Van Zoest, R.A., Leech, R., Wit, F.W., Portegies, P., Geurtsen, G.J., Schmand, B.A. et al. Increased brain-predicted aging in treated HIV disease. *Neurology*, 88(14):1349–1357, apr 2017.

[21] Van Schependom, J., Vidaurre, D., Costers, L., Sjøgård, M., D'hooghe, M.B., D'haeseleer, M., Wens, V., De Tiège, X., Goldman, S., Woolrich, M. et al. Altered transient brain dynamics in multiple sclerosis: Treatment or pathology? *Human Brain Mapping*, 40(16):4789–4800, nov 2019.

[22] Costers, L., Van Schependom, J., Laton, J., Baijot, J., Sjøgård, M., Wens, V., De Tiège, X., Goldman, S., D'Haeseleer, M., D'hooghe, M.B. et al. Spatiotemporal and spectral dynamics of multi-item working memory as revealed by the n-back task using MEG. *Human Brain Mapping*, 41(9):2431–2446, jun 2020.

[23] Benedict, R.H., Deluca, J., Phillips, G., LaRocca, N., Hudson, L.D. and Rudick, R. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis, apr 2017.

[24] Van Schependom, J., D'hooghe, M.B., Cleynhens, K., D'hooge, M., Haelewyck, M., De Keyser, J. and Nagels, G. The Symbol Digit Modalities Test as sentinel test for cognitive impairment in multiple sclerosis. *European Journal of Neurology*, 21(9):1219–e72, sep 2014.

[25] Sandry, J., Simonet, D.V., Brandstadter, R., Krieger, S., Katz Sand, I., Graney, R.A., Buchanan, A.V., Lall, S. and Sumowski, J.F. The Symbol Digit Modalities Test (SDMT) is sensitive but non-specific in MS: Lexical access speed, memory, and information processing speed independently contribute to SDMT performance. *Multiple Sclerosis and Related Disorders*, 51, jun 2021.

[26] Jain, S., Sima, D.M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., De Mey, J., Barkhof, F., Steenwijk, M.D., Daams, M. et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage: Clinical*, 8:367–375, jun 2015.

[27] Le, T.T., Kuplicki, R.T., McKinney, B.A., Yeh, H.W., Thompson, W.K. and Paulus, M.P. A Nonlinear Simulation Framework Supports Adjusting for Age When Analyzing BrainAGE. *Frontiers in Aging Neuroscience*, 10, oct 2018.

[28] Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E. and Miller, K.L. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, oct 2019.

[29] Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C. and Mechelli, A. Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*, 72(103600):1–9, oct 2021.

[30] Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R. and Kievit, R.A. Raincloud plots: A multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. *Wellcome Open Research*, 4:63, apr 2019.

[31] Golan, D., Doniger, G.M., Srinivasan, J., Sima, D.M., Zarif, M., Bumstead, B., Buhse, M., Van Hecke, W., Wilken, J. and Gudesblatt, M. The association between MRI brain volumes and computerized cognitive scores of people with multiple sclerosis. *Brain and Cognition*, 145:105614, nov 2020.

[32] Uher, T., Vaneckova, M., Sormani, M.P., Krasensky, J., Sobisek, L., Dusankova, J.B., Seidl, Z., Havrdova, E., Kalincik, T., Benedict, R.H. 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, feb 2017.

[33] Macaron, G., Baldassari, L.E., Nakamura, K., Rao, S.M., McGinley, M.P., Moss, B.P., Li, H., Miller, D.M., Jones, S.E., Bermel, R.A. et al. Cognitive processing speed in multiple sclerosis clinical practice: association with patient-reported outcomes, employment and magnetic resonance imaging metrics. *European Journal of Neurology*, 27(7):1238–1249, jul 2020.

[34] Ribeiro, M.T., Singh, S. and Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1135–1144, aug 2016.

[35] Wood, D.A., Kafiabadi, S., Busaidi, A.A., Guilhem, E., Montvila, A., Lynch, J., Townend, M., Agarwal, S., Mazumder, A., Barker, G.J. et al. Accurate brain-age models for routine clinical MRI examinations. *NeuroImage*, 249:118871, apr 2022.

[36] Benedict, R.H., Weinstock-Guttman, B., Fishman, I., Sharma, J., Tjoa, C.W. and Bakshi, R. Prediction of Neuropsychological Impairment in Multiple Sclerosis: Comparison of Conventional Magnetic Resonance Imaging Measures of Atrophy and Lesion Burden. *Archives of Neurology*, 61(2):226–230, feb 2004.

[37] D'hooghe, M.B., Gielen, J., Van Remoortel, A., D'haeseleer, M., Peeters, E., Cambron, M., De Keyser, J. and Nagels, G. Single MRI-Based Volumetric Assessment in Clinical Practice Is Associated With MS-Related Disability. *Journal of Magnetic Resonance Imaging*, 49(5):1312–1321, may 2019.

[38] Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[39] Peters, R. Ageing and the brain, feb 2006.

[40] Ruigrok, A.N., Salimi-Khorshidi, G., Lai, M.C., Baron-Cohen, S., Lombardo, M.V., Tait, R.J. and Suckling, J. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39:34–50, 2014.

[41] Bethlehem, R.A.I., Seidlitz, J., White, S.R., Vogel, J.W., Anderson, K.M., Adamson, C., Adler, S., Alexopoulos, G.S., Anagnostou, E., Areces-Gonzalez, A. et al. Brain charts for the human lifespan. *Nature 2022 604:7906*, 604(7906):525–533, apr 2022.

[42] Nobis, L., Manohar, S.G., Smith, S.M., Alfaro-Almagro, F., Jenkinson, M., Mackay, C.E. and Husain, M. Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical*, 23:101904, 2019.

[43] Rajkomar, A., Dean, J. and Kohane, I. Machine Learning in Medicine. *N Engl J Med*, 380(26):2589–2590, jun 2019.

[44] Shapiro, S.S. and Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[45] Jiang, H., Lu, N., Chen, K., Yao, L., Li, K., Zhang, J. and Guo, X. Predicting Brain Age of Healthy Adults Based on Structural MRI Parcellation Using Convolutional Neural Networks. *Frontiers in Neurology*, 10:1346, jan 2020.

[46] Wittens, M.M.J., Allemeersch, G.J., Sima, D.M., Naeyaert, M., Vanderhasselt, T., Vanbinst, A.M., Buls, N., De Brucker, Y., Raeymaekers, H., Fransen, E. et al. Inter- and Intra-Scanner Variability of Automated Brain Volumetry on Three Magnetic Resonance Imaging Systems in Alzheimer's Disease and Controls. *Frontiers in Aging Neuroscience*, 13, oct 2021.

[47] Lysandropoulos, A.P., Absil, J., Metens, T., Mavroudakis, N., Guisset, F., Van Vlierberghe, E., Smeets, D., David, P., Maertens, A. and Van Hecke, W. Quantifying brain volumes for Multiple Sclerosis patients follow-up in clinical practice – comparison of 1.5 and 3 Tesla magnetic resonance imaging. *Brain and Behavior*, 6(2):e00422, feb 2016.

[48] Poole, M.A. and O'Farrell, P.N. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, pages 145–158, 1971.

[49] Denissen, S., Engemann, D.A., De Cock, A., Costers, L., Baijot, J., Laton, J., Penner, I.K., Grothe, M., Kirsch, M., D'hooghe, M. et al. Brain age as a surrogate marker for information processing speed in multiple sclerosis. *MedRxiv*, pages 2021–09, 2021.

[50] Khalil, M., Teunissen, C., Langkammer, C. et al. Iron and neurodegeneration in multiple sclerosis. *Multiple sclerosis international*, 2011, 2011.

[51] Hanke, M., Baumgartner, F.J., Ibe, P., Kaule, F.R., Pollmann, S., Speck, O., Zinke, W. and Stadler, J. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, 1(1):1–18, may 2014.

[52] Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C. and Buckner, R.L. Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, sep 2007.

[53] Nooner, K.B., Colcombe, S.J., Tobe, R.H., Mennes, M., Benedict, M.M., Moreno, A.L., Panek, L.J., Brown, S., Zavitz, S.T., Li, Q. et al. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in Neuroscience*, 6(OCT):152, oct 2012.

[54] Landman, B.A., Huang, A.J., Gifford, A., Vikram, D.S., Lim, I.A.L., Farrell, J.A., Bogovic, J.A., Hua, J., Chen, M., Jarso, S. et al.

Multi-parametric neuroimaging reproducibility: A 3-T resource study. *NeuroImage*, 54(4):2854–2866, feb 2011.

[55] IXI dataset. http://brain-development.org/ixi-dataset/.

[56] Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L. and Petersen, S.E. Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84:320–341, jan 2014.

[57] Vidorreta, M., Wang, Z., Rodriguez, I., Pastor, M.A., Detre, J.A. and Fernández-Seara, M.A. Comparison of 2D and 3D single-shot ASL perfusion fMRI sequences. *NeuroImage*, 66:662–671, feb 2013.

[58] Van Schuerbeek, P., Baeken, C. and De Mey, J. The Heterogeneity in Retrieved Relations between the Personality Trait 'Harm Avoidance' and Gray Matter Volumes Due to Variations in the VBM and ROI Labeling Processing Settings. *PLOS ONE*, 11(4):e0153865, apr 2016.

[59] Clare Kelly, A.M., Uddin, L.Q., Biswal, B.B., Castellanos, F.X. and Milham, M.P. Competition between functional brain networks mediates behavioral variability. *NeuroImage*, 39(1):527–537, jan 2008.

[60] Chen, M.Y., White, C.N., Giles, N., Elumn, A., Parikh, S., Kim, U., Maddox, W.T. and Poldrack, R.A. OpenNeuro Dataset ds000053 (Training of loss aversion modulates neural sensitivity toward potential gains.), 2017.

[61] Maas, A.I., Menon, D.K., David Adelson, P.D., Andelic, N., Bell, M.J., Belli, A., Bragge, P., Brazinova, A., Büki, A., Chesnut, R.M. et al. Traumatic brain injury: Integrated approaches to improve prevention, clinical care, and research, dec 2017.

[62] Papadopoulos Orfanos, D., Michel, V., Schwartz, Y., Pinel, P., Moreno, A., Le Bihan, D. and Frouin, V. The Brainomics/Localizer database. *NeuroImage*, 144:309–314, jan 2017.

[63] Diedenhofen, B. and Musch, J. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, 10(4):e0121945, apr 2015.

[64] Pearson, Karl and Filon. VII. Mathematical contributions to the theory of evolution.— IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical*

Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 191:229–311, dec 1898.

Chapter 9

Federated learning for brain image decoding in multiple sclerosis

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], 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

Under review
This chapter is based on a preprint on *medRxiv* [1]

Abstract

Introduction: Deep learning research requires lots of centralized data. However, data sets are often stored at different clinical centres, and sharing sensitive patient data such as brain images is difficult. In this manuscript, we investigated the feasibility of federated learning (FL) for research on brain magnetic resonant images of people with multiple sclerosis (MS).

Methods: Four computers were connected in the same virtual private network. 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). Using this network, we fine-tuned a previously published brain age model to decode performance on the symbol digit modalities test (SDMT) of patients with MS from structural T1 weighted brain MRI. Model training happened with the previously published FedAvg algorithm, sending models and results via secure copy protocol.

Results: Training consisted of 22 federated learning rounds. The resulting 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.

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

9.1 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 [2, 3], and can already be considered indispensable less than 50 years after Peter Mansfield successfully scanned the finger of his assistant Andrew A. Maudsley [4].

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 [5]. 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" [6]. It is plausible that such representations should be enriched with other biological information, such as functional brain organisation [7]. However, besides resolving to other methodologies, recent evidence suggests that more information can be extracted from structural MRI than common knowledge-based representations [8].

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 [8]. 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 [9]. Analogously to Leonardsen et al. 2022 [8], we will now use transfer 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

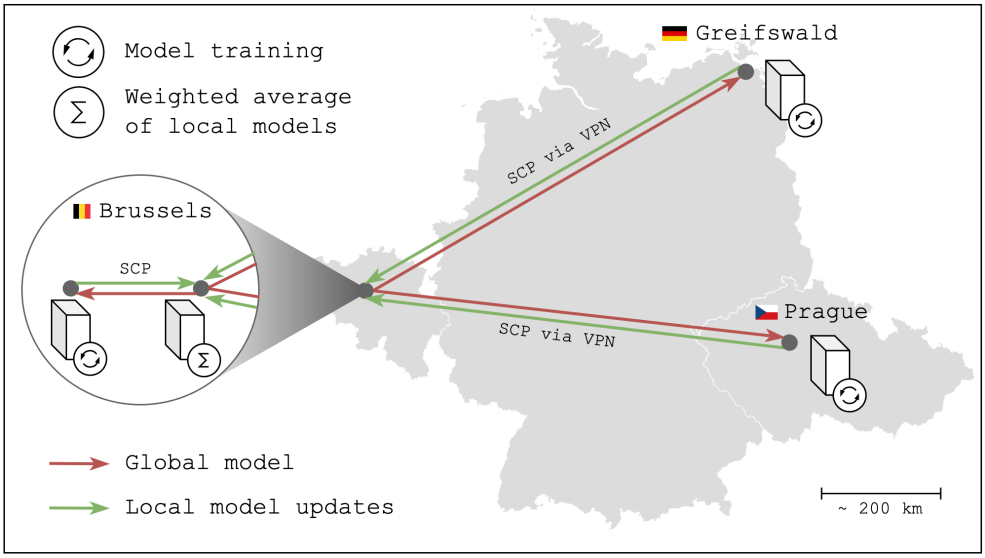

Figure 9.1: The federated learning network. The computer with the "sigma" symbol is the server, whereas computers with an "update" symbol are clients. Abbreviations: SCP = Secure Copy Protocol; VPN = Virtual Private Network.

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) [10], 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 [11]. The feasibility of federated learning was underlined by a recent study on brain tumour segmentation using a world-wide network of 71 sites [12].

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.

9.2 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 centre. For each centre in the federated learning network (figure 9.1), T1 weighted MR images were available, as well as demographic and clinical information. This entailed sex, age, expanded disability status scale (EDSS [13], overall disability), symbol digit modalities test (SDMT [14], explained below), disease duration and MS subtype. Preprocessing of T1 weighted MR images was performed using the preprocessing pipeline of Wood et al. 2022 [15], 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 [16] (v1.13.1) since v1.7.1 did not work with Compute Unified Device Architecture (CUDA) v12.1 [17]. Data were organised locally in the Brain Imaging Data Structure (BIDS) format [18] to facilitate decentralised model training and evaluation. Data are described in table 9.1. 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.

Brain age model

We used a pre-trained T1 brain age model from Wood et al. 2022 [15], which the authors made available in their GitHub repository. This model was chosen for three reasons: (1) it is a deep neural network that only uses brain images as input, (2) it has a state-of-the-art, low error in predicting age from structural MRI and (3) their methodology could be replicated using their code, from

	Brussels	Greifswald	Prague	p value
n	97	104	100	
sex (m:f)	28:69	35:69	24:76	$0.315^{\dagger}$
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	$0.018^{\dagger}$

Table 9.1: 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 Kallner et al. 2017 [19] and was used to avoid data sharing.

data preprocessing to predicting brain age. In the context of this study, we used the deep learning model as a feature extractor. This yields a data-driven latent representation of 1024 features (cfr. figure 9.2), which might retain more information from the original image compared to the knowledge-based feature space of chapter 8, containing only 12 volumetric features.

The model is a Dense Convolutional Network (DenseNet) [20], which was already shown to outperform other network architectures in a medical imaging context [21]. A DenseNet is unique for directly connecting all layers inside a "dense block" with each other [20]. Each layer therefore takes all previous feature maps, outputs of previous layers, as input. Hence, the propagation of features throughout the network is improved. The DenseNet architecture moreover reduces the "vanishing gradient" problem, where the gradient used to update weights in the network gradually approaches zero during backpropagation to earlier layers. Lastly, it reduces the number of parameters in the network. The 3D DenseNet used in this study has a total of 11243649 parameters (weights and biases), and is characterised by 4 dense blocks, consisting respectively of 6, 12, 24 and 16 dense layers. Although the exact model architecture can be consulted in the paper of Wood et al. 2022 [15], in the context of transfer learning in this manuscript (cfr. next section), it is noteworthy to mention the size of the fully connected layer, consisting of 1024 weights and 1 bias (figure 9.2).

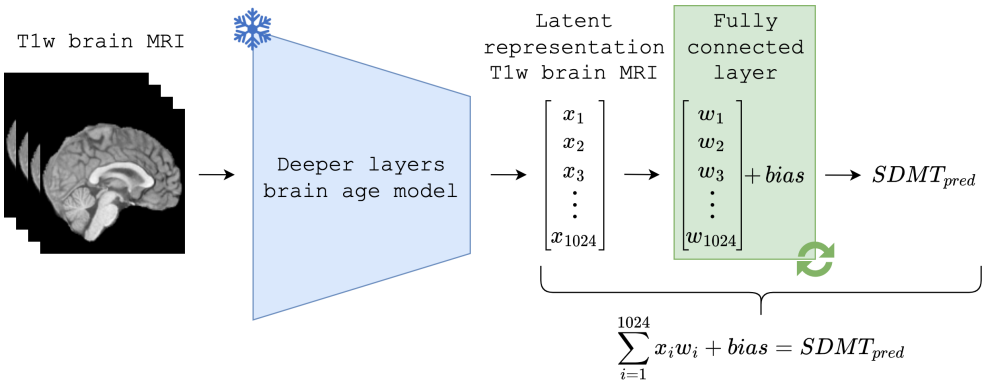

Figure 9.2: Transfer learning methodology. The deeper layers of the 3D DenseNet were frozen during training, whereas the fully connected layer (including 1024 weights and 1 bias) was updated. In between the deeper layers and the fully connected layer is the latent (data-driven) representation of a T1w brain MRI.

Transfer learning

To update the brain age model to predict SDMT with transfer learning, we used the "Feature extractor" approach discussed in Kim et al. 2022 [22]. In this approach, the original fully connected layer is used, which in this case is a linear regression with 1024 independent variables (the latent features). Only the weights of this layer are updated; the weights of all deeper layers are frozen, i.e., they are not updated during training. This approach was chosen in light of the size of the decentralised data set, analogous to Leonardsen et al. 2022 [8].

Age decoding performance

First, we applied the brain age model of Wood et al. 2022 [15] to data of 50 healthy controls from the Brussels client to establish the generalisability 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 [9]. Furthermore, as brain age models typically overestimate age of MS patients [9, 23], 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.

Hardware setup

The federated learning network (figure 9.1) 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 [10]. 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 [24]. This FL plan contains all details for local model training and can be consulted in appendix A. 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 9.2), 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:

$$MAE_{test,overall} = \sum_{i=0}^{m} \frac{MAE_{test,i} * n_i}{N}$$

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 centre 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.

9.3 Results

Brain age predictions

The brain age model of Wood et al. 2022 [15] achieved an MAE of 3.85 years on the Brussels HC data set, whereon it significantly underestimated age (table 9.2). The model overestimated age on the Brussels and Greifswald data set (table 9.2). BPAD distributions of the client MS data sets were significantly different (p <.001, calculated with an ANOVA on n, mean and SD of table 9.2).

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

Federated learning

Figure 9.3 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. 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.

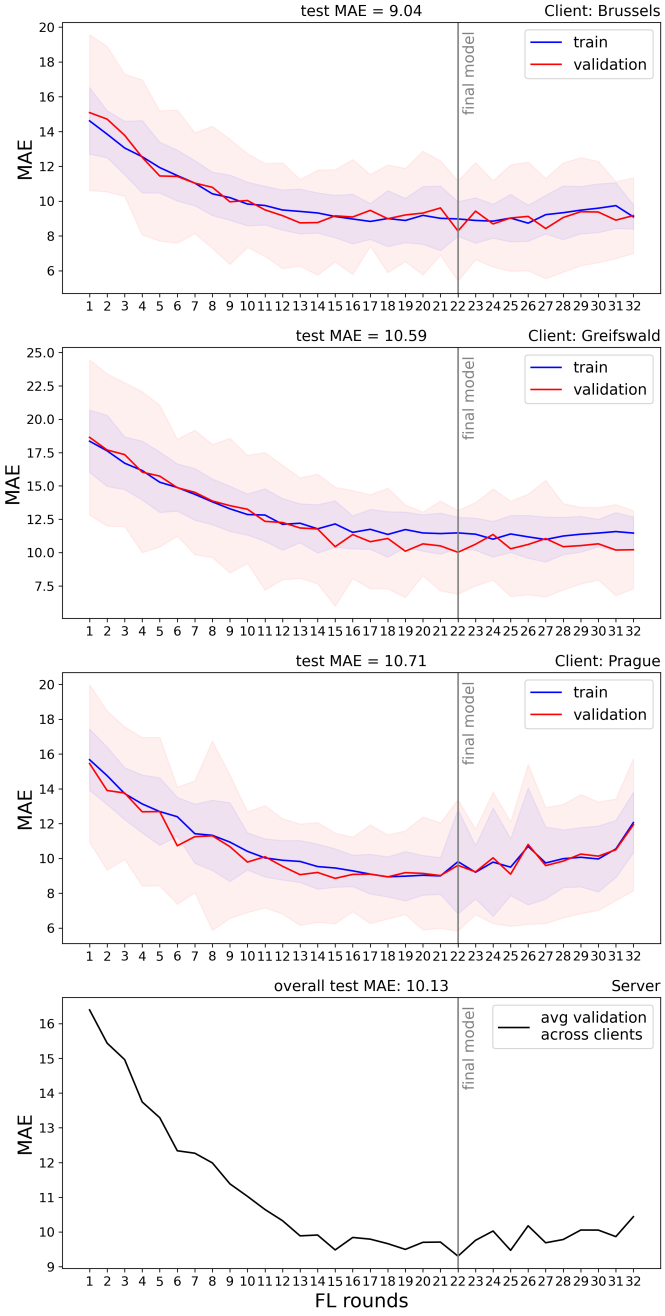

Figure 9.3: Federated learning results. Abbreviations: FL = federated learning, MAE = mean absolute error, avg = average. Final model: FL round 22.

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 9.3.

		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 9.3: 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.

9.4 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 centres. On all three clients, the performance in decoding SDMT from T1 weighted brain images gradually improved, resulting in a final federated learning model 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). Respectively for Brussels, Greifswald and Prague, the Pearson correlation between true and predicted SDMT was 0.30 ($p = 0.206$), 0.29 ($p = 0.210$) and 0.54 ($p = 0.014$).

SDMT decoding performance of the FL model

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 generally 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 9.4).

The key observation for this table is the standard deviation of the predicted SDMT distribution, which is 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. If 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. Several factors could explain this observation:

	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 9.4: 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.

- **Data heterogeneity.** Although we harmonised the MRI data using the preprocessing pipeline of Wood et al. 2022, the client data sets differ on various demographical and clinical features (cfr. table 9.1). These differences might explain the lower test MAE in the Brussels data set compared to the other data sets, and the fact that the test MAE differs more between clients in the client-specific model training with respect to the federated learning approach. Although the MR images were preprocessed prior to training, data harmonisation on other sample characteristics could have improved the performance across clients. For true generalisability to be obtained, however, the model should also have acceptable performance on cases with different characteristics. This problem might be circumvented by adopting a larger decentralised database (cfr. infra).
- **Sample size.** This proof-of-concept study modelled on about 100 cases per client. Increasing the decentralised database with more data per client, or by including more clients in the network, might boost the performance of the network. We furthermore hypothesise that this allows using other transfer learning methodologies, as discussed next.

- **Transfer learning methodology.** We might have frozen too many network weights, thereby overestimating the similarity of the age and SDMT decoding task. Although MS is characterised by neurodegenerative processes that also occur in healthy aging, which in turn partly explains their cognitive performance [1], the T1 weighted image might reflect MS-related damage that is not represented in the latent space of the brain age model. With increased sample size, other transfer learning methodologies might be considered such as unfreezing more layers, or replacing the fully connected layer with another regressor, such as a random forest regression [22]. This former allows making the latent features more specific for the SDMT decoding task, while the latter allows non-linear interactions between the latent features and SDMT to be learned.

  Choosing another loss function, for example L2 loss instead of L1 loss, might also impact model performance. In L2 loss, the error between true and predicted label is squared, thereby more severely penalising extreme mispredictions compared to reasonably accurate predictions. As this boosts the model's incentive to avoid large errors, the tendency of learning to assign values close to the mean might reduce, in turn improving individual predictions.

- **T1 weighted brain MRI.** Even when unfreezing all layers of the network, a T1-weighted brain MR image might not contain sufficient information to decode the information processing speed of a subject with MS, as it mostly captures the neurodegenerative aspect of the disease. Neuroinflammatory damage is visible to some extent as black holes in T1 weighted brain images, but is better visualised by other image modalities such as FLuid-Attenuated Inversion Recovery (FLAIR) images. As lesion volume is important in the prediction of cognitive impairment [25], a combination of T1-weighted and FLAIR brain MRI could increase the performance of the model. The characteristic neuroinflammatory damage caused by MS however does not occur in a healthy ageing cohort, implying that a brain age model might be insensitive for this.

Neuroscientific implications

The fact that the model only performed poorly limited further exploration of the model. When addressing the aforementioned issues, we hypothesise that the performance of the SDMT-decoding model can be significantly improved to accepted clinical standards, such as decoding the SDMT performance within

the clinically meaningful change of 4 points [26]. Beyond methodological insights that were presented in this study, this could yield new neuroscientific insights in the clinico-radiological paradox. In skin cancer research, for example, examining a skin cancer prediction model with explainable AI (XAI) guided the focus of human experts to the background of the images, which is commonly overlooked. The authors hypothesise the underlying cause to be "visual entrenchment" [27]. In a similar way, we hypothesise that by examining an accurate SDMT decoding model, novel insights in the structural underpinnings of cognitive impairment could be obtained.

The federated learning approach

Our approach to federated learning can be considered basic, but its simplicity makes it transparent. 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 UNiplexed Information Computing System (UNIX)-based operating systems.

Currently, open source federated learning frameworks are in full development, such as Flower [28], OpenFL [24] and PySyft [29]. 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. Federated learning in MS is still in its infancy, but promising to ultimately boost AI research in MS.

9.5 Conclusion

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

9.6 Code availability

In order to support future FL projects, we made our code publicly available in the GitHub repository of our lab, the Artificial Intelligence-supported Modelling in clinical Sciences (AIMS) lab of the Vrije Universiteit Brussel (VUB). Link: https://github.com/AIMS-VUB/FLightcase.

9.7 Appendix: Federated learning 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). The upper bound of the number of FL rounds was set to 100 in light of training time. Adapting this number turned out not to be necessary as model convergence consistently happened before this number was reached.
- Number of epochs per FL round: 1. One epoch is one complete model update on all available training data. To simplify the training process, we fixed this to 1 epoch, meaning that every FL round, the model was trained only once on the entire training dataset. In this way, the number of FL rounds is equivalent to the number of epochs in a centralised setting.
- Batch size: 8. Number of data points used simultaneously to calculate the gradient, which allows to update all weights in a model simultaneously. We chose a batch size of 8 as it is common in our domain [30], and in light of the memory capacity (12GB) of the graphical processing unit of the Brussels node.
- 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. In the Wood et al. 2022 paper [15], 0.0001 was used, but we chose a larger learning rate to speed up learning. Both learning rates are common in deep learning research in medical imaging and yield acceptable results [31, 32].
- 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). We reduced this number with respect to Wood et al. 2022 (the authors used 5 [15]) to act earlier when learning stagnates.
- 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. Instead of a train/validation/test split of 65/15/20 in Wood et al. 2022 [15], we used 60/20/20 to increase the number of samples in the validation data set.
- 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. McMahan et al. 2017 suggests to calculate a weighted average across a subsample of the total number of clients [10]. With a total of 3 clients in the sample, we therefore chose to average across 2 clients, as we deemed a contribution of a single client to be suboptimal to take model variability into account.
- Number of splits: 30. Number of random train/validation splits (using bootstrapping) for each FL round. The number was chosen arbitrarily, but in light of a trade-off between sufficient splits relative to the number of cases and training time.
- Loss function: L1 loss, $\sum_{i=1}^{n} |y_i - \hat{y}|$. Summed absolute error between true and predicted SDMT score. We used the L1 loss as it is closely related to the mean absolute error ($MAE = \frac{L1\_loss}{n}$), which was used for its intuitive interpretation as the average points of SDMT misprediction.
- Optimiser: To update the weights, we used Adam optimisation. This method has several interesting properties, such as fast convergence [33].

References

[1] Denissen, S., Grothe, M., Vaneckova, M., Uher, T., Laton, J., Kudrna, M., Horakova, D., Kirsch, M., Motyl, J., De Vos, M. et al. Transfer learning on structural brain age models to decode cognition in MS: a federated learning approach. *medRxiv*, 2023.

[2] Kaunzner, U.W. and Gauthier, S.A. 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] Traboulsee, A.L. and Li, D. The role of MRI in the diagnosis of multiple sclerosis. *Advances in neurology*, 98:125–146, 2006.

[4] Mansfield, P. and Maudsley, A.A. Medical imaging by NMR. *The British journal of radiology*, 50(591):188–194, 1977.

[5] Wattjes, M.P., Rovira, À., Miller, D., Yousry, T.A., Sormani, M.P., De Stefano, N., Tintore, M., Auger, C., Tur, C., Filippi, M. 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.

[6] Barkhof, F. The clinico-radiological paradox in multiple sclerosis revisited. *Current opinion in neurology*, 15(3):239–245, 2002.

[7] Sjøgård, M., Wens, V., Van Schependom, J., Costers, L., D'hooghe, M., D'haeseleer, M., Woolrich, M., Goldman, S., Nagels, G. and De Tiège, X. Brain dysconnectivity relates to disability and cognitive impairment in multiple sclerosis. *Human brain mapping*, 42(3):626–643, 2021.

[8] Leonardsen, E.H., Peng, H., Kaufmann, T., Agartz, I., Andreassen, O.A., Celius, E.G., Espeseth, T., Harbo, H.F., Høgestøl, E.A., de Lange, A.M. et al. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage*, 256:119210, 2022.

[9] Denissen, S., Engemann, D.A., De Cock, A., Costers, L., Baijot, J., Laton, J., Penner, I.K., Grothe, M., Kirsch, M., D'hooghe, M.B. et al. Brain age as a surrogate marker for cognitive performance in multiple sclerosis. *European Journal of Neurology*, 29(10):3039–3049, 2022.

[10] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A. Communication-efficient learning of deep networks from decentralized

data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[11] Sheller, M.J., Reina, G.A., Edwards, B., Martin, J. and Bakas, S. 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.

[12] Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.H., Reina, G.A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C. et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):1–17, 2022.

[13] Kurtzke, J.F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1444, 1983.

[14] Smith, A. *Symbol digit modalities test*. Western psychological services Los Angeles, 1973.

[15] Wood, D.A., Kafiabadi, S., Al Busaidi, A., Guilhem, E., Montvila, A., Lynch, J., Townend, M., Agarwal, S., Mazumder, A., Barker, G.J. et al. Accurate brain-age models for routine clinical MRI examinations. *Neuroimage*, 249:118871, 2022.

[16] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[17] Nickolls, J., Buck, I., Garland, M. and Skadron, K. Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for? *Queue*, 6(2):40–53, 2008.

[18] Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.

[19] Kallner, A. Resolution of Students t-tests, ANOVA and analysis of variance components from intermediary data. *Biochemia medica*, 27(2):253–258, 2017.

[20] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[21] Zeng, W., Peng, J., Wang, S., Li, Z., Liu, Q. and Liang, D. A comparative study of CNN-based super-Resolution methods in MRI reconstruction. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1678–1682. IEEE, 2019.

[22] Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E. and Ganslandt, T. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

[23] Kaufmann, T., van der Meer, D., Doan, N.T., Schwarz, E., Lund, M.J., Agartz, I., Alnæs, D., Barch, D.M., Baur-Streubel, R., Bertolino, A. et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623, 2019.

[24] Foley, P., Sheller, M.J., Edwards, B., Pati, S., Riviera, W., Sharma, M., Moorthy, P.N., Wang, S.h., Martin, J., Mirhaji, P. et al. OpenFL: the open federated learning library. *Physics in Medicine & Biology*, 67(21):214001, 2022.

[25] Uher, T., Vaneckova, M., Sormani, M., Krasensky, J., Sobisek, L., Dusankova, J.B., Seidl, Z., Havrdova, E., Kalincik, T., Benedict, R. 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.

[26] Benedict, R.H., Deluca, J., Phillips, G., LaRocca, N., Hudson, L.D. and Rudick, R. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis, apr 2017.

[27] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J. et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine 2020 26:8*, 26(8):1229–1234, jun 2020.

[28] Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., de Gusmão, P.P. and Lane, N.D. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

[29] Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D. and Passerat-Palmbach, J. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017*, 2018.

[30] Anand, V., Gupta, S., Altameem, A., Nayak, S.R., Poonia, R.C. and Saudagar, A.K.J. An enhanced transfer learning based classification for diagnosis of skin cancer. *Diagnostics*, 12(7):1628, 2022.

[31] Hinton, B., Ma, L., Mahmoudzadeh, A.P., Malkov, S., Fan, B., Greenwood, H., Joe, B., Lee, V., Kerlikowske, K. and Shepherd, J. Deep learning networks find unique mammographic differences in previous negative mammograms between interval and screen-detected cancers: a case-case study. *Cancer imaging*, 19(1):1–9, 2019.

[32] Li, J., Wang, P., Zhou, Y., Liang, H. and Luan, K. Different machine learning and deep learning methods for the classification of colorectal cancer lymph node metastasis images. *Frontiers in Bioengineering and Biotechnology*, 8:620257, 2021.

[33] Kingma, D.P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Part III

The future of AI in MS

Chapter 10

Will artificial intelligence change MS care within the next 10 years?

Artificial intelligence is a hot topic nowadays. OpenAI's ChatGPT was released in late 2022 [1], and quickly reached world-wide popularity as chat bot, providing highly accurate answers to complicated questions and even producing software code. Despite the plethora of envisioned use cases, ChatGPT has been subjected to a lot of criticism as well; some claim to embrace it as a tool to catalyse research and development [2], while others underline the threats to for example the originality of scientific work [3]. This illustrates the complexity of the AI debate, which is no different in medical contexts. How will AI shape our future?

At the end of 2022, Multiple Sclerosis Journal published a series of papers that concern the short-term future perspective of AI for MS care. The section in the journal is termed "Controversies in Multiple Sclerosis", and contains one paper that defends a statement, one that attacks a statement, and the last providing a commentary. The statement that we treated in the series was "Artificial intelligence will change MS care within the next 10 years". My promoter, Prof. Guy Nagels, and I defended the statement [4], while my promoter Prof. Jeroen Van Schependom and Prof. Maarten De Vos attacked the statement [5]. Two researchers of the University of Campania Luigi Vanvitelli commented on the discussion [6].

As discussed earlier in chapter 4, few AI models reach clinical practice in MS. Why would this be any different in the next 10 years? The arguments made by the defending paper [4] can be consulted in the paper that is in-

cluded at the end of this chapter. In summary, AI already supports clinical workflow in MS by providing quantitative summaries of MR images. We can however expect more, since medicine appears to be ready to welcome AI. This is for example illustrated by emerging real-world examples that AI increases diagnostic accuracy, that AI is increasingly included in medical training and that humans and AI are increasingly collaborating. Personalised treatment is however not yet a reality, and one of the reasons for this might be in the data that is used for modelling, for example lacking quality or quantity.

The attacking paper [5] claims that the prosper of AI in MS care is dependent on the data used to train them. Large data sets are necessary, although it is unclear how large they need to be. If databases are to be enlarged by combining data of multiple clinical centres, which is difficult in light of the General Data Protection Regulation (GDPR), the data should be harmonised. Moreover, data continuously change due to advances in clinical practice, making old data less useful for modelling purposes. Data is also key in assessing the generalisability of a model, as data that is used only for testing purposes might subtly leak to the training data, giving the model an unfair advantage. Lastly, the authors highlight the difficulty in bringing AI models to the clinic. Even if the performance of a model is satisfactory as proven by a clinical trial, the model is likely a complicated one of which the working is poorly understood, potentially causing clinicians to not trust the model.

The overall conclusion of the authors that commented on both papers was that AI might in the future assist MS neurologists, but not within 10 years [6].

Artificial intelligence will change MS care within the next 10 years: Yes

Stijn Denissen[1,2], Guy Nagels[1,2,3]

1 AIMS Lab, Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium **2** icometrix, Leuven, Belgium **3** St Edmund Hall, University of Oxford, Queen's Lane, Oxford, UK

This chapter is based on a paper in *Multiple Sclerosis Journal* [4]

Artificial intelligence (AI) changes our experience in daily life. Every day, AI improves our digital consumer experience by offering personalised advertisements, our global communication by powering accurate translation machines and helps us find our way through a plethora of information on the World Wide Web. An ever-digitalizing world with increasing complexity needs AI as a guiding compass.

10.1 AI supports medicine

This is also true for medicine. In an acute life-threatening situation such as stroke, AI is guiding personalised treatment decisions that must be made in a split second to reduce further brain damage. [7] During colonoscopy procedures, a real-time AI-powered device, already approved for clinical use in the European Union and United States, reduces the number of missed colorectal neoplasia cases by half. [8] AI also reduces workload and costs in actual clinical settings like infection detection. [9] These examples show that in medical fields outside of multiple sclerosis (MS), AI already improves diagnostic performance, workflow, and cost-effectiveness.

10.2 AI is in full development

Several developments in the field of AI indicate that we will see many more AI-based algorithms in clinical practice in the upcoming 10 years. First, explainable AI (XAI) methods are in full development and help clinicians trust AI models. This is important because technological evolution, for example, deep learning, makes AI models more performant at the cost of increased complexity. Explanations can be retrieved even from complex neural networks, and besides increasing our understanding on how such networks come to their decisions, they could guide a diagnostic process. [10] For example, by revealing the regions that a convolutional neural network (CNN) deemed important for the diagnosis of actinic keratoses, medical students were taught to also take the background of an image into account, which the authors hypothesise to be otherwise often overlooked. [10]

A second development is formed by newly developed training courses, for instance, training radiology registrars in AI. [11] These meet clinicians' need for additional training to understand both strengths and weaknesses of AI tools, not unlike their need for basic statistical knowledge to implement evidence-based medicine. This need is addressed in a third new perspective, namely the shift from a human–computer competition viewpoint (can AI do better than a human doctor?) to a human–computer collaboration. [10]

Finally, AI research groups also increasingly realise that the evaluation of AI tools should not exclusively happen in laboratory situations but should also be executed 'under real-world conditions in the hands of the intended users'. [10] An AI system trained and tested on real-world screening mammograms resulted in improved breast cancer screening and generalised to mammograms of patients from the United States when trained only on mammograms of patients from the United Kingdom. [12]

10.3 AI supports workflow in MS care

Recent AI developments also support routine work within clinical workflows. An example is that quantification of brain magnetic resonance imaging (MRI) volumes by Food and Drug Administration (FDA)-approved AI-based segmentation algorithms already allows us to assess MRI disease activity in MS in a quantitative way, facilitating clinical use of modern criteria of therapeutic efficacy such as the 'No Evidence of Disease Activity' (NEDA) 3 and 4 criteria.

10.4 AI impacts treatment planning

Apart from facilitating the often complicated clinical workflow for MS, AI will help us with other unmet needs in MS care. An important clinical question is always 'how will my patient progress in the future, and how can I positively influence this course by choosing the right therapy?' Personalised treatment planning is already feasible in an acute stroke setting, 1 but is still under development in the MS setting. To reach this goal, we need more and better data, and improved AI models. With regard to data, the AI boom motivates clinicians to contribute to existing international data collections such as MSBase, coordinated from Australia, and it leverages large international government-funded projects for data-interoperability such as EHDEN (European Health Data & Evidence Network), coordinated by the Department of Medical Informatics in Erasmus MC in the Netherlands. Such data collections have already resulted in prognostic models achieving good accuracy in predicting mortality in oncology. [13] AI tools for image quantification have also encouraged neuroradiologists to adopt three-dimensional, instead of two-dimensional, T1 and FLAIR (fluid-attenuated inversion recovery) images into standard practice for MS patients, further increasing data quality.

10.5 AI drives novel drug development

Finally, personalised treatment planning should not be static, in the sense that we should not be restricted to the currently existing and mostly immune-mediated therapies. Rather, we need to find and prescribe treatments to our patients that would also protect them against neurodegeneration, that would ultimately halt MS and improve the damage caused by the disease. This seems like a very distant target, but AI can also help us speed up the discovery process for these new therapies, thereby extending our therapeutic arsenal.

First, AI could help in clinical trial design. This was already shown in glaucoma where an AI model successfully identified high-risk patients [14], which leads to shorter study duration and/or lower patient numbers needed to meet clinical trial endpoints. Second, by improving screening results of compound libraries, AI methods have already been producing tangible results in the development of new antibiotics for over a decade. [15] As MS neuroprotection is a sorely needed therapeutic area, it is encouraging to read that recently AI methods were used to identify a sirtuin-1 active compound, for which the neuroprotective and pro-regenerative effect was subsequently confirmed in a

sciatic nerve crush animal model. [16]

10.6 Conclusion

AI notably positively influences medical practice in different fields. Several developments in methods, clinician awareness, and data quality/quantity are aligning to also create this impact in the MS field within the next few years.

References

[1] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z. et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.

[2] Van Dis, E.A., Bollen, J., Zuidema, W., van Rooij, R. and Bockting, C.L. ChatGPT: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[3] Thorp, H.H. ChatGPT is fun, but not an author, 2023.

[4] Denissen, S. and Nagels, G. Artificial intelligence will change MS care within the next 10 years: Yes. *Multiple Sclerosis Journal*, 28(14):2171–2173, 2022.

[5] De Vos, M. and Van Schependom, J. Artificial intelligence will change MS care within the next 10 years: No. *Multiple Sclerosis Journal*, 28(14):2173–2174, 2022.

[6] Lavorgna, L. and Bonavita, S. Artificial intelligence will change MS care within the next 10 years: Commentary. *Multiple Sclerosis Journal*, 28(14):2175–2176, 2022.

[7] Grunwald, I.Q., Ragoschke-Schumm, A., Kettner, M., Schwindling, L., Roumia, S., Helwig, S., Manitz, M., Walter, S., Yilmaz, U., Greveson, E. et al. First Automated Stroke Imaging Evaluation via Electronic Alberta Stroke Program Early CT Score in a Mobile Stroke Unit. *Cerebrovascular Diseases*, 42(5-6):332–338, nov 2016.

[8] Wallace, M.B., Sharma, P., Bhandari, P., East, J., Antonelli, G., Lorenzetti, R., Vieth, M., Speranza, I., Spadaccini, M., Desai, M. et al. Impact of Artificial Intelligence on Miss Rate of Colorectal Neoplasia. *Gastroenterology*, 163(1):295–304.e5, jul 2022.

[9] Burton, R.J., Albur, M., Eberl, M. and Cuff, S.M. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Medical Informatics and Decision Making*, 19(1):1–11, aug 2019.

[10] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J. et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine 2020 26:8*, 26(8):1229–1234, jun 2020.

[11] Lindqwister, A.L., Hassanpour, S., Lewis, P.J. and Sin, J.M. AI-RADS: An Artificial Intelligence Curriculum for Residents. *Academic Radiology*, 28(12):1810–1816, dec 2021.

[12] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A. et al. International evaluation of an AI system for breast cancer screening. *Nature 2020 577:7788*, 577(7788):89–94, jan 2020.

[13] Morin, O., Vallières, M., Braunstein, S., Ginart, J.B., Upadhaya, T., Woodruff, H.C., Zwanenburg, A., Chatterjee, A., Villanueva-Meyer, J.E., Valdes, G. et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nature Cancer 2021 2:7*, 2(7):709–722, jul 2021.

[14] Chen, A., Montesano, G., Lu, R., Lee, C.S., Crabb, D.P. and Lee, A.Y. Visual field endpoints for neuroprotective trials: a case for AI driven patient enrichment. *American Journal of Ophthalmology*, 2022.

[15] Urbina, F., Puhl, A.C. and Ekins, S. Recent advances in drug repurposing using machine learning. *Current Opinion in Chemical Biology*, 65:74–84, dec 2021.

[16] Romeo-Guitart, D., Forés, J., Herrando-Grabulosa, M., Valls, R., Leiva-Rodríguez, T., Galea, E., González-Pérez, F., Navarro, X., Petegnief, V., Bosch, A. et al. Neuroprotective Drug for Nerve Trauma Revealed Using Artificial Intelligence. *Scientific Reports*, 8(1):1879–1879, jan 2018.

Chapter 11

Discussion and future work

The central aim of this thesis was to use AI to obtain new insights in the relationship between structural brain imaging and cognitive performance in people with MS. To construct AI models, sufficient data needs to be present, especially for deep learning research. Data availability however was the key limiting factor, as large open source cognition-labelled imaging databases are not available for MS, and data sharing between clinical centres is difficult, for example due to privacy considerations.

11.1 Three solutions for limited data availability

This thesis explored three solutions for limited data availability. First, ico**gnition**, a smartphone-based cognitive assessment, was introduced to facilitate digital data collection and creation of research databases. The other two solutions are AI techniques that respectively reduce the need for data (transfer learning) or enhance the accessibility of data without data sharing (federated learning).

Solution 1: Digital data collection

Medicine in general experiences a shift towards digitalisation. This has many benefits such as practising medicine remotely (telemedicine) and storing data digitally. The latter facilitates the creation of research databases that can subsequently be used to train AI models. In chapter 7, the validity and reliability of ico**gnition** was assessed, a smartphone-based cognitive screening battery for people with MS. The tests in the cognitive screening battery correlated well with their paper-pencil equivalents, correlated with other clinical variables and had a moderate-to-good test-retest reliability. The latter indic-

ates the stability of test performance over time when cognitive performance is expected to be similar. Surprisingly, the test performance of people with MS and healthy control subjects was similar. We hypothesise to have recruited an MS sample that was relatively spared in cognitive performance.

Conclusion 1: ico**gnition** is a valid and reliable smartphone-based cognitive screening battery for people with MS.

Solution 2: Reducing the need for data

Transfer learning uses the similarity of the task of interest with another task for which more data is available. A robust model is trained to perform the related task, which is then fine-tuned to solve the task of interest using less data. Here, the task of interest is the prediction of cognition from structural brain MRI from people with MS. A related task that can be used is the prediction of age from structural brain MRI, for which many open source data sets are available. A first step, however, is to prove the similarity between both tasks. This similarity was shown in chapter 8; brain age correlates with cognitive performance in people with MS.

Conclusion 2: Older looking brains are associated with worse cognition in people with MS.

In chapter 9, the actual transfer learning was carried out, fine-tuning a brain age model to a cognition model. The results appeared promising at first. The mean absolute error (the average number of points SDMT misprediction) decreased by updating the model on the data set of each clinical centre using federated learning (cfr. infra). By looking at the distribution of the SDMT predictions however, we discovered that the model probably used a "trick" to lower the error: it appeared to start assigning values close to the mean. This could result in a lower L1 loss (summed absolute error) when accurate individual predictions cannot be made. This does however not mean that we can immediately label brain age as a useless intermediary step towards a cognition model. Other methodological choices might lead to different results. Some methodological choices are discussed below.

1. **Model architecture.** In chapter 8, a simple linear regression model (12 weights) was used to predict age from a feature representation of a brain image. Linear regression is the simplest form of a neural network which comes with a great advantage: it is very easy to understand. The other extreme is a deep learning network such as the one used in chapter 9, consisting of millions of weights. Combined with the high-dimensional

input (pixel space), this model is accurate at the cost of explainability. This trade-off between performance and explainability should be taken into account when designing the model architecture.

2. **How many layers to freeze.** For the transfer learning approach, only the last layer of the brain age model was updated during training. This means that we are assuming that the brain age latent space representation of a brain MR image contains sufficient information to predict cognition. By allowing deeper layers of the models to be updated, the similarity between both tasks (predicting age and cognition) becomes gradually less important, while the necessity for a larger training data set increases. The number of layers to freeze however appears to be determined through experimentation, analogously to many other methodological considerations in AI research such as model architecture [1].

3. **Other transfer learning methods.** Besides varying the number of layers to freeze, Kim et al. 2022 describe another method in which the machine learning model after the feature extractor is replaced [1]. Another machine learning model might be better suited to predict the label (e.g. SDMT performance) from the latent variables.

4. **Image type.** Brain age models are typically trained on T1-weighted brain MR images, although exceptions exist (e.g. Wood et al. 2022 used T2-weighted images [2]). An important consideration here is the availability of data. An age-labelled database of T1-weighted MR images is relatively easy to construct since they are abundantly available in open source data repositories. Modelling on T1 images will allow sensitising models for neurodegeneration, but not the inflammatory activity in the CNS. However, even if other image modalities that highlight inflammatory activity (e.g. FLAIR images) were to be used, MS brain lesions do not occur in healthy controls. In future studies, it would be very interesting to investigate whether there is a performance saturation when fine-tuning multimodal (multiple image types) brain age models to decode cognition.

Conclusion 3: Transfer learning does not allow training a brain age model to predict cognition. Different transfer learning parameters such as the number of weights to freeze might improve performance in future studies.

Solution 3: Enhancing data accessibility

In federated learning (FL), a model is trained by sending models instead of data. The key idea is that models are not trained at one central place. Instead, models are trained locally where data is present, after which they are collected and processed at one central place. The interesting property of FL is that data is not shared, but the data is still accessible for machine learning research. In chapter 9, an international FL network was constructed and its feasibility proven by fine-tuning a brain age model to a cognition model.

Conclusion 4: Federated learning is feasible for training models in cognitive neuroscience research in MS in a decentralised way.

Future avenues for FL in a medical context are described next.

11.2 What's next in federated learning?

In chapter 9, the feasibility of a relatively simple approach to federated learning was shown. By sending models via secure copy protocol between 4 computers in 3 countries, present in a virtual private network, consistent learning behaviour of a neural network was observed. There are however several considerations to take into account for future endeavours.

11.2.1 Privacy

The bottom line of why federated learning was invented in the first place, is to circumvent issues related to sharing data. One important consideration is maintaining the privacy of the individual whose data is used for modelling, i.e. the data subject [3]. But even when data is not shared, privacy concerns are still present in deep learning research, such as the "Indirect (Inferred) Information Exposure" in Mireshghallah et al. 2020 [4], where data can be inferred from e.g. properties of a model. There are multiple ways of quantifying the loss of privacy, of which one is "Differential Privacy" [5]. Here, a data holder/curator promises a data subject the following: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available." [3]. Several techniques are proposed that aim to fulfil this promise, such as adding random noise to the data being modelled upon [6]. Several existing open source toolboxes implement differential privacy, such as Opacus [7] and PySyft [8]. Both use a method called Differentially Private Stochastic Gradient Descent (DP-SGD) [9]. It is explained both narratively and visually

in Yousefpour et al. 2021 [7] (Opacus). A gradient is computed per sample, after which the L2 norm is clipped. Subsequently, a batch gradient is made by aggregating the sample gradients, after which Gaussian noise is added [7].

Another way of addressing privacy concerns in AI is through the use of generative AI. It generates synthetic data, that could mitigate modelling on real patient data when sufficiently realistic [10]. A popular example of a generative AI model is a generative adversarial network (GAN), consisting of a generator and a discriminator, which can be regarded as a villain and a police officer respectively [11]. For brain MR images for example, the villain tries to generate realistic, fake brain images, while the police tries to discriminate fake from real images. By challenging each other, both gradually increase performance, training the generator to produce more realistic brain images.

11.2.2 How to investigate model performance in a decentralised way?

One of the problems we faced when performing federated learning was the evaluation of the final model after training it. Since data cannot be shared, the true ground truth labels (in our case SDMT performance) should also be kept locally. In the study described in chapter 9, each client computer shared descriptors of both the predicted and true ground truth distributions with the server, which could then be analysed. Although the data indicated that the model did not provide meaningful individual predictions, it is impossible to thoroughly analyse the behaviour of the model on an individual level.

The question therefore arises whether ground truth labels and predictions can be shared with the server to allow thorough model evaluation. Scatter plots have been around for decades to evaluate the relationship between two variables, which visually and anonymously reveals the individuals values of a certain variable. Federated learning is a young research field, and future research should reach consensus about guidelines in model assessment in a decentralised context.

11.2.3 Methodological considerations

The most popular way of performing federated learning is the federated averaging (FedAvg) algorithm proposed by McMahan et al. 2016 [12]. This is however not the only method for aggregating models, which are outlined chronologically in Moshawrab et al. 2023 [13]. The privacy concerns outlined

above are also reflected in the aggregation process, for example the "Differential Privacy Average Aggregation" method [13], including a "privacy budget" parameter to control the level of privacy conservation during aggregation.

11.2.4 The contribution of clinical partners

Federated learning could provide a unique way of driving international collaboration between clinical partners. Clinical input for model development is indispensable for models to be explainable and trustworthy, while AI-related research questions can be developed in a multidisciplinary setting. This assures model robustness from a technical point of view, while maximising clinical relevance of the research questions being addressed.

11.3 The role of XAI in future AI studies

Explainable AI was introduced in chapter 4, but has received little attention in this thesis beyond the discussion on user trust in chapter 8. In the context of investigating the relationship between structural MRI and cognition in MS, the central theme of this thesis, XAI could however play an important role in the future. Assuming future studies are able to train more accurate models to predict cognitive performance from structural MRI when sufficient high-quality data is available, these models can be assessed with XAI. Brain regions that the model deemed important for the prediction can subsequently be highlighted in the input images. This additional information allows clinical MS experts to compare the model's behaviour with their expertise in the field, potentially inducing trust in the model [14]. Moreover, it could lead to new insights in the relationship between structural MRI and cognition in MS.

11.4 The big picture and future perspective

The ultimate goal of any research endeavour in MS is to improve the well-being of patients. As we live in an increasingly digitalising era, it is important to take advantage of available tools such as smartphones and high computational power to generate and process data towards better understanding and management of the disease. Especially in a heterogeneous disease like MS, personalised medicine is crucial to optimise care. Although this thesis did not offer a solution for the paradox between structural damage on brain images and cognitive impairment, it showcased how we can make better use of currently available digital tools to address the paradox and many other research

questions still to come. In terms of cognitive follow-up for example, training AI models on ico**gnition** data could personalise cognitive follow-up by identifying patient-specific trends and predict potential deterioration.

Linear regression was a central methodological choice in this thesis. It was first used to construct a brain age model, and later on in the context of transfer learning to obtain a cognition decoding model. This might be an oversimplification from a neurobiological point of view, but it is in general considered good practice to start with a simple benchmark model, on which can be improved subsequently. Indeed, in the context of brain age, linear regression for example ignores the fact that white matter and hippocampal volume increases until approximately middle age (chapter 8). The same argument of oversimplification applies to (1) using only T1-weighted brain images to model the ageing brain and cognitive impairment, on which neuroinflammatory damage is less visible, and (2) only training the last layer of a brain age network to predict cognitive performance.

Besides the argument of benchmarking, methodological choices were limited by data availability. By presenting three solutions for this problem in this thesis, future research might be able to explore more complicated models that better approximate the disease process occurring in MS. In this light, it is important to keep in mind that more complex models come at the cost of reduced interpretability. We might therefore risk using a model of which we are unaware of limitations and biases. Explainable AI could offer a solution in this regard, and might additionally reveal new insights in the clinico-radiological paradox, much like XAI helped guide human focus on the background of skin images for diagnosing skin cancer [15]. It should moreover be highlighted that the results presented in this study are not limited to MS. Especially the concept of transfer learning and federated learning are universally applicable, while after validation, ico**gnition** might prove useful in other diseases characterised by cognitive decline such as Alzheimer's Disease.

In conclusion, this thesis presented benchmark results of using digital tools to address the clinico-radiological paradox in MS. It underlines the importance to invest in targeting a sufficiently large data base, driving meaningful AI research. By addressing this important issue in the future, AI could help demystifying the clinico-radiological paradox in MS, ultimately leading to informed clinical decision making in addressing the burdensome symptom that is cognitive impairment in MS.

References

[1] Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E. and Ganslandt, T. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

[2] Wood, D.A., Kafiabadi, S., Busaidi, A.A., Guilhem, E., Montvila, A., Lynch, J., Townend, M., Agarwal, S., Mazumder, A., Barker, G.J. et al. Accurate brain-age models for routine clinical MRI examinations. *NeuroImage*, 249:118871, apr 2022.

[3] Dwork, C., Roth, A. et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[4] Mireshghallah, F., Taram, M., Vepakomma, P., Singh, A., Raskar, R. and Esmaeilzadeh, H. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*, 2020.

[5] Qin, S., He, J., Fang, C. and Lam, J. Differential private discrete noise adding mechanism: Conditions, properties and optimization. *arXiv preprint arXiv:2203.10323*, 2022.

[6] He, J. and Cai, L. Differential private noise adding mechanism and its application on consensus. *arXiv preprint arXiv:1611.08936*, 2016.

[7] Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J. et al. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[8] Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D. and Passerat-Palmbach, J. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017*, 2018.

[9] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[10] Murdoch, B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(1):1–5, 2021.

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[13] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H. and Raad, A. Reviewing Federated Learning Aggregation Algorithms; Strategies, Contributions, Limitations and Future Perspectives. *Electronics*, 12(10):2287, 2023.

[14] Diprose, W.K., Buist, N., Hua, N., Thurier, Q., Shand, G. and Robinson, R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, 27(4):592–600, 2020.

[15] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J. et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine 2020 26:8*, 26(8):1229–1234, jun 2020.

Curriculum Vitae

Stijn Denissen was born in Haaren, the Netherlands on the 31st of May 1994. After finishing secondary school at Gymnasium Beekvliet in Sint-Michielsgestel, the Netherlands in 2012 (combined track Nature & Health and Nature & Technology), he moved to Leuven, Belgium to study rehabilitation sciences at the KU Leuven. He obtained his bachelor in rehabilitation sciences in 2016, and his master in rehabilitation sciences (track neurorehabilitation) with distinction in 2018. His master thesis, in collaboration with Anne Lubbe, was entitled "Trunk rehabilitation in the different recovery phases post-stroke: a systematic review and meta-analysis" (promotor: Prof. Dr. Geert Verheyden).

After meeting Prof. Dr. ir. Guy Nagels at the NMSC Melsbroek in 2018, he started his PhD in medical sciences at the Vrije Universiteit Brussel in 2019, for which he obtained a Baekeland grant from Flanders Innovation and Entrepreneurship (VLAIO) in December 2019.

Publications

First author

- Denissen, S., Engemann, D.A., De Cock, A., Costers, L., Baijot, J., Laton, J., Penner, I.K., Grothe, M., Kirsch, M., D'hooghe, M.B., D'Haeseleer, M., Dive, D., De Mey, J., Van Schependom, J., Sima, D.M. and Nagels, G., 2022. Brain age as a surrogate marker for cognitive performance in multiple sclerosis. *European journal of neurology*, 29(10), pp.3039-3049
- Denissen, S. and Nagels, G., 2022. Artificial intelligence will change MS care within the next 10 years: Yes. *Multiple Sclerosis Journal*, 28(14), pp.2171-2173
- Denissen, S., Chén, O.Y., De Mey, J., De Vos, M., Van Schependom, J., Sima, D.M. and Nagels, G., 2021. Towards multimodal machine learning

prediction of individual cognitive evolution in multiple sclerosis. *Journal of Personalized Medicine*, 11(12), p.1349

- Denissen, S., De Cock, A., Meurrens, T., Vleugels, L., Van Remoortel, A., Gebara, B., D'Haeseleer, M., D'Hooghe, M.B., Van Schependom, J. and Nagels, G., 2019. The impact of cognitive dysfunction on locomotor rehabilitation potential in multiple sclerosis. *Journal of central nervous system disease*, 11, p.1179573519884041

- Denissen, S., Staring, W., Kunkel, D., Pickering, R.M., Lennon, S., Geurts, A.C., Weerdesteyn, V. and Verheyden, G.S., 2020. Interventions for preventing falls in people after stroke. *Stroke*, 51(3), pp.e47-e48

- Denissen, S., Staring, W., Kunkel, D., Pickering, R.M., Lennon, S., Geurts, A.C., Weerdesteyn, V. and Verheyden, G.S., 2019. Interventions for preventing falls in people after stroke. *Cochrane database of systematic reviews*, (10)

Shared first author

- Van Laethem, D., Denissen, S., Costers, L., Descamps, A., Baijot, J., Van Remoortel, A., Van Merhaegen-Wieleman, A., D'Hooghe, M.B., D'Haeseleer, M., Smeets, D., Sima, D.M., Van Schependom, J., and Nagels, G., 2023. The Finger Dexterity Test: validation study of a smartphone-based manual dexterity assessment. *Multiple Sclerosis Journal*, 13524585231216007

- Baijot, J., Denissen, S., Costers, L., Gielen, J., Cambron, M., D'Haeseleer, M., D'hooghe, M.B., Vanbinst, A.M., De Mey, J., Nagels, G. and Van Schependom, J., 2021. Signal quality as Achilles' heel of graph theory in functional magnetic resonance imaging in multiple sclerosis. *Scientific Reports*, 11(1), p.7376

Co-author

- Thijs, L., Voets, E., Denissen, S., Mehrholz, J., Elsner, B., Lemmens, R. and Verheyden, G.S., 2023. Trunk training following stroke. *Stroke*, 54(9), pp.e427-e428

- Thijs, L., Voets, E., Denissen, S., Mehrholz, J., Elsner, B., Lemmens, R. and Verheyden, G.S., 2023. Trunk training following stroke. *Cochrane Database of Systematic Reviews*, (3)

- Baijot, J., Van Laethem, D., Denissen, S., Costers, L., Cambron, M., D'Haeseleer, M., D'hooghe, M.B., Vanbinst, A.M., De Mey, J., Nagels, G. and Van Schependom, J., 2022. Radial diffusivity reflects general decline rather than specific cognitive deterioration in multiple sclerosis. *Scientific Reports*, 12(1), p.21771
- De Keersmaecker, E., Beckwée, D., Denissen, S., Nagels, G. and Swinnen, E., 2021. Virtual reality for multiple sclerosis rehabilitation. *Cochrane Database of Systematic Reviews*, 2020(12), p.CD013834

Preprints

- Denissen, S., Van Laethem, D., Baijot, J., Costers, L., Descamps, A., Van Remoortel, A., Van Merhaegen-Wieleman, A., D'hooghe, M.B., D'Haeseleer, M., Smeets, D., Sima, D.M., Van Schependom, J., and Nagels, G., 2023. icognition: a smartphone-based cognitive screening battery. *medRxiv*, pp.2023-07
- Denissen, S., Grothe, M., Vaneckova, M., Uher, T., Laton, J., Kudrna, M., Horakova, D., Kirsch, M., Motyl, J., De Vos, M., Chen, O.Y., Van Schependom, J., Sima, D.M. and Nagels, G., 2023. Transfer learning on structural brain age models to decode cognition in MS: a federated learning approach. *medRxiv*, pp.2023-04

Grants

- Baekeland grant (HBC.2019.2579) by Flanders Innovation and Entrepreneurship. Industrial PhD grant, collaboration Vrije Universiteit Brussel and ico**metrix**.
- Fonds Wetenschappelijk Onderzoek (FWO) travel grant (V412023N) for a research stay in Prague, Czech Republic.
- European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS) travel grant for the ECTRIMS 2023 conference in Milan.

Presentations

International conferences

- Rehabilitation in multiple sclerosis (RIMS) conference 2019, Ljubljana, Slovenia

  - Talk 1: "The impact of cognitive dysfunction on locomotor rehabilitation potential in Multiple Sclerosis"
  - Talk 2: "The effects of multidisciplinary rehabilitation in Multiple Sclerosis"

- Neurorehabilitation and Neural Repair (NNR) conference 2019, Maastricht, the Netherlands. Talk: "Interventions for preventing falls in people after stroke"

- Americas Committee for Treatment and Research in Multiple Sclerosis (ACTRIMS) Forum 2021, virtual conference. Talk: "Brain age in MS: An explainable principal component of brain MRI and potential sensitive cognitive biomarker"

- OpenMR Virtual 2021, virtual conference

  - Workshop: "Introduction to data exploration with Python, Matlab/Octave and Jupyter Notebooks"
  - Talk: "A paper in a song"
  - Moderator on panel discussion about science communication

- International Multiple Sclerosis Cognition Society (IMSCOGS) 2022, Bordeaux, France. Talk: "Brain age as a surrogate marker for IPS in MS"

Regional/national talks

- Research talks VUB: international mobility, 2023. Recording experience stay abroad in Prague + discussion with vice rector research

- National MS Centrum Melsbroek symposium "Wetenschappelijk Onderzoek in het NMSC: projectresultaten en toepassingen", Elewijt, 2023. Talk: "Icognition & Finger Dexterity Test - test je cognitieve functie & fijne motoriek op je smartphone"

- Lustrum (20 years) symposium of physical therapy science, Utrecht University. Invited panel member. Question addressed: "Will artificial intelligence reshape future physiotherapy science?".

- VUB book presentation "Truth", Brussels city hall, 2023. Interviewee.

- 2nd MS Nurse academy 2019, Brussels, Belgium. Talk: "Clinical data registration"

- VUB PhD day 2023. Live performance of "a paper in a song" + interviewee on PhD experience at VUB
- Award ceremony "Vlaamse scriptieprijs 2022", Brussels city hall, 2023. Live performance of "a paper in a song" and other science-related songs.
- C4N PhD day 2019. "Music as instrument for science communication" (live performance of "a paper in a song")

Science communication

- Chapter in the VUB book "Truth": "Medical diagnosis and a new kid in town: Artificial Intelligence"
- Workshop "Hoe oud is mijn brein?", VUB children's university, 2021.
- Science blog for "Wtnschp"
  - "Hersenspinsels: een editie over cognitie", 2021
  - "Hersenspinsels: artificiële intelligentie", 2020
- YouTube channel "a paper in a song", converting a paper into a song with animation. Presented live at several occasions (cfr. supra)

Figure 11.1: QR code to the "a paper in a song" YouTube channel

- Article in Wetenshappelijk Onderzoek in MS (WOMS), scientific magazine of MS-Liga Vlaanderen: "Artificiële intelligentie (AI) voor het voorspellen van de evolutie bij MS"

Media

- Article in "Haelio Neurology" by Julia Ernst, MS. Title: "Brain age serves as sensitive marker for information processing speed in MS"

- Article in "Artsenkrant" by Dr. Michèle Langendries. Title: "De hersenleeftijd als bevattelijke klinische parameter bij MS"

Board member

- Natural Sciences and Engineering (NSE) PhD network
- OpenMR Virtual 2021 conference

Conference posters

- ECTRIMS/ACTRIMS joint meeting 2023, Milan, Italy. Poster title: "Transfer learning on structural brain age models to decode cognition in MS: a federated learning approach"
- ECTRIMS/ACTRIMS joint meeting 2020, virtual conference. Poster title: "Predicted brain age as a cognitive biomarker in multiple sclerosis". Published abstract in Multiple Sclerosis Journal: Denissen, S., De Cock, A., Costers, L., Baijot, J., Laton, J., D'Hooghe, M. B., D'Haeseleer, M., Dive, D., De Mey, J., Van Schependom, J., Sima, D. M., & Nagels, G. (2020). Predicted brain age as a cognitive biomarker in multiple sclerosis. Multiple Sclerosis Journal, 26(S3), 124-125. [P0015].
- Computational approaches for ageing and age-related diseases (CompAge) 2020, virtual conference. Poster title: "Predicted brain age as a cognitive biomarker in multiple sclerosis"
- C4N PhD day 2019. Poster title: "The impact of cognitive dysfunction on locomotor rehabilitation potential in multiple sclerosis"
- Rehabilitation in multiple sclerosis (RIMS) conference 2019, Ljubljana, Slovenia
  - Poster title 1: "The impact of cognitive dysfunction on locomotor rehabilitation potential in Multiple Sclerosis"
    Published abstract in Multiple Sclerosis Journal: Denissen, S., De Cock, A., Meurrens, T., Vleugels, L., Van Remoortel, A., Gebara, B., D'Haeseleer, M., D'Hooghe, M. B., Van Schependom, J., & Nagels, G. (2019). The Impact of Cognitive Dysfunction on Locomotor Rehabilitation Potential in Multiple Sclerosis. Multiple Sclerosis Journal, 25(7), 1041-1041.

  - Poster title 2: "The effects of multidisciplinary rehabilitation in Multiple Sclerosis"
    Published abstract in Multiple Sclerosis Journal: Denissen, S., Noë, S., Ferdinand, S., Vleugels, L., Gebara, B., Nagels, G., & Meurrens, T. (2019). The Effects of Multidisciplinary Rehabilitation in Multiple Sclerosis. Multiple Sclerosis Journal, 25(7), 1040-1041.

- Neurorehabilitation and Neural Repair (NNR) conference 2019, Maastricht, the Netherlands. Poster title: "Interventions for preventing falls in people after stroke"

Index

R

S

T

U

V