UCLouvain

École polytechnique de Louvain

Towards the integration of advanced magnetic resonance imaging biomarkers in multiple sclerosis clinical practice

Author: **Maxence WYNEN**
Supervisors: **Benoit MACQ, Pietro MAGGI**
Readers: **Meritxell BACH CUADRA, Eliott BRION, Michel VERLEYSEN**
Academic year 2020–2021
Master [120] in Data Sciences Engineering

Remerciements / Acknowledgements

Ce mémoire est le l'aboutissement d'un énorme investissement personnel, marquant la fin d'un cursus mémorable à l'École Polytechnique de Louvain. Je ne serais cependant pas arrivé au même résultat sans l'aide de quelques personnes.

Je souhaite tout d'abord remercier mes promoteurs: Pr Benoît Macq pour ses conseils avisés et sa confiance; et Dr Pietro Maggi, qui au terme d'une année passée à collaborer avec lui a su me transmettre sa passion pour la recherche en sclérose en plaques.

I would also like to thank Dr Meritxell Bach Cuadra and Francesco La Rosa for their precious help and technical support even from another country. I am glad to be joining you next year in Lausanne.

Enfin, j'aimerais remercier mes amis et ma famille pour leur soutien et leurs encouragements inconditionnels tout au long de l'écriture de ce mémoire.

Abstract

Multiple sclerosis (MS) is a chronic immune-mediated disease characterized by focal inflammatory lesions in the brain and spinal cord. Magnetic Resonance Imaging (MRI) is an essential tool in the diagnosis of MS. However, markers currently used in the clinic are not specific enough for MS. Several advanced MRI biomarkers, including paramagnetic rim lesions (PRL), more specific to MS and based on advanced MRI techniques have recently been studied.

This thesis aims to i) develop a multi-modal database including clinical, biological and imaging features of MS patients; ii) explore the heterogeneity of the disease using statistics and machine learning techniques; iii) to contribute to the collective effort of researchers in the field to automate the extraction of MRI features and iv) to introduce the issue of longitudinal detection of advanced MRI features in MS.

The first part of this thesis focuses on the completion of the multi-modal database. It first presents the tools developed to automate the MRI features' extraction: a pipeline for converting from the DICOM format to the BIDS standard as well as a pipeline for the extraction of volumetric data and the localization of lesions. Then, this work shows that RimNet, a 3D U-Net inspired Convolutional Neural Network for automated PRL detection, has a good generalization capacity and that it is able to make longitudinal consistent predictions. This part ends with the specification of a database management system specific to the context along with a proposal of solution.

The second part presents the results of statistical analysis, and describes in detail how our database resembles or not the current literature in the field. It also announces the steps still in progress to perfect the objectives, i.e. the techniques of imputation of data to be explored, the regression for the prognosis and the clustering for the detection of the potential subtypes of MS.

The impact of this thesis is manifold. First, it has created an innovative database, including clinical, biological and both established and advanced imaging features. On the way to the completion of this database, several tools have been developed which have also made it possible to speed up the work of researchers in this field by automating specific time-consuming tasks. The thesis also lays the foundations for the development of a multi-modal database management system in the medical imaging sector. Finally, the preliminary analysis of this multi-modal database, already helped to get more insights on the heterogeneus nature of the disease in different MS patients.

Contents

List of Figures

Chapter 1

Introduction

Multiple sclerosis (MS) is a chronic immune-mediated disease of the brain and spinal cord (central nervous system - CNS), affecting about 2 million people worldwide with a prevalence of about 1 per 1000 in Northern Europe. Magnetic resonance imaging (MRI) is the established method to observe in vivo the pathological processes occurring in the CNS of MS patients, and is thus considered an essential tool to diagnose and to follow/predict the evolution of the disease (disease prognosis). However, the specificity of the established diagnostic criteria for MS (mainly based on MRI) is limited. This may result in misdiagnosing MS in patients affected by other illnesses and in a limited accuracy in predicting conversion to MS in "at risk" individuals. Improving diagnostic accuracy is a major unmet clinical-research need in MS.

On the other hand, due to the low pathological specificity of the MRI techniques currently used in the clinical routine, the pathological heterogeneity of what is diagnosed as “MS” is probably underestimated. Indeed MS is a heterogeneous condition, probably a “spectrum disorder”, regrouping many different diseases sharing clinical, biological and imaging features.[1]

To meet this demand, MS-specific biomarkers based on advanced MRI features have recently been studied. These include (but are not limited to) the central vein sign (CVS) [2][3][4], paramagnetic rim lesion (PRL) biomarker [5][6], intra or juxta-cortical lesions [7] and brain atrophy [8].

This thesis is part of a larger multi-disciplinary project aiming to i) improve the differentiation between MS and MS mimicking disorders, ii) identify different pathological entities with the "MS spectrum" and iii) better predict conversion to MS in "at risk" patients. To achieve these goals, a very important step was to create and fill an innovative multi-modal database including clinical, biological and both classical and advanced MRI features. With such a unique database, we hope to discover insights not only on the MS differential diagnosis and intra-disease heterogeneity (by classification/clustering), but also on MS prognosis (e.g. by regression). This can be done by classical Machine Learning.

Additionally, the main issue with the filling of such a database is that the assessment of these advanced MRI features is currently done manually by medical

experts, which is very time consuming and prone to intra-rater and inter-raters variability. To avoid this, several automated feature extraction methods have been developed [9][10] and this thesis reports a preliminary attempt to longitudinal detect these MRI features using Machine Learning [11]. Overall, the long-term goal of this research is to obtain a fully automated pipeline in which the MRI features are automatically detected and combined with as a final goal taking into account the most relevant clinical and MRI features to speed-up and improve both the diagnosis and the prognosis of MS. Finally, this thesis shows that the collected data are in line with the available scientific evidences on this subject, [6][5][3][12][13][14] thus supporting the methods we choose to fill-up this thesis database.

In short, the objectives of this thesis are:

- To develop a multi-modal database including clinical, biological and imaging features of MS patients
- To explore the MS disease heterogeneity using statistical analysis and machine learning
- To contribute to the researcher's effort to automate classic and advanced MRI feature extraction
- To introduce the problematic of longitudinal detection of MRI features in MS

Chapter 2

Theoretical Background

2.1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is an imaging technique invented fairly recently (second half of the 20^{th} century) by two teams, lead by Felix Bloch (Stanford) and Edward M. Purcell (Harvard). The technology used by MRI allows the acquisition of a variety of images (anatomical, functional, etc.) of biological tissues in a non-invasive way. The underlying physics principles of MRI make it one of the safest, if not the safest, imaging technique available. This whole section, greatly inspired by [15], aims to give a quick overview of the physics behind how MRI creates images.

2.1.1 Nuclear magnetism

Every atom's nucleus is made up of a certain number of protons and neutrons (nucleons) animated by a complex collective movement including an individual rotation around an axis going through their own center. Every particle spinning around itself has an angular momentum or "spin" aligned on its rotation axis (represented by a vector $\overrightarrow{S}$). Moreover, every spinning charge induces a magnetic field around itself: this *magnetic moment* is represented by the vector $\overrightarrow{\mu}$ (Figure 2.2).

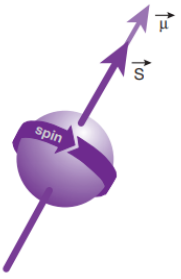


Figure 2.1: Spin of a particle. $\overrightarrow{S}$: axis of rotations; $\overrightarrow{\mu}$: magnetic moment. Source: [15]

Given that all nucleons are made up of positive and negative sub-particles (called

quarks, Figure 2.2), they also possess a magnetic moment although having a neutral electrical charge. In fact, every nucleon is constituted of three quarks linked together. Quarks can be *up* (electrical charge of $+2/3$) or *down* (electrical charge of $-1/3$). In particular, protons are made up of two quarks *up* and one quark *down* (and thus an electrical charge of $+2/3 + 2/3 - 1/3 = +1$) while neutrons are made up of two quarks *down* and one quark *up* (electrical charge of $-1/3 - 1/3 + 2/3 = 0$). Therefore, only the atoms with an uneven number of nucleons possess a total magnetic moment.

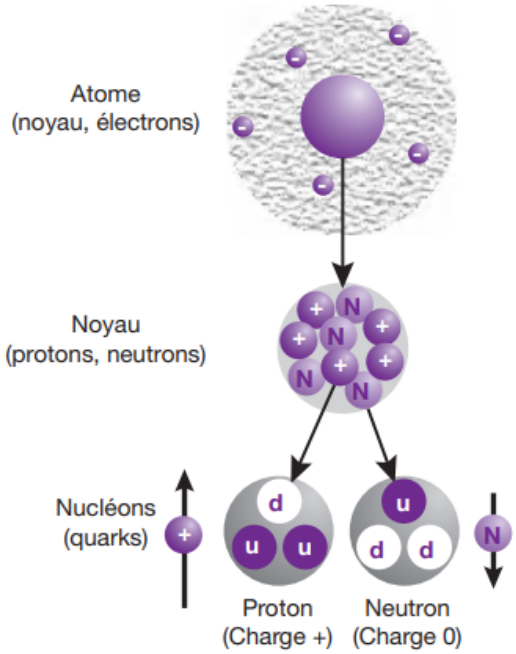


Figure 2.2: Structure of the atom. Source: [15]

The only nuclei of biological interest possessing magnetic properties we will consider in MRI are the nuclei of hydrogen (^{1}H), because hydrogen is estimated to constitute two-thirds of the organism's atoms. These protons (hydrogen nuclei) are thus assimilated to small magnets (magnetic dipoles) with a magnetic moment $\overrightarrow{\mu}$. The net macroscopic magnetization vector (Figure 2.3) is defined by the sum of all these magnetic moments:

$$\overrightarrow{M} = \sum_i \overrightarrow{\mu_i}$$

2.1.2 Magnetic resonance

Protons are placed in a large (generally 1.5T, 3T or 7T) external magnetic field $\overrightarrow{B_0}$ tend to orient themselves in its direction. In particular, protons subdivide into two sub-populations rotating at a certain angle around the z axis (precession): one in the same direction as $\overrightarrow{B_0}$ (parallel) and the other one in the exact opposite direction (anti-parallel). It is important to note that the ratio $\frac{N_{parallel}}{N_{anti-parallel}}$ (that depends on

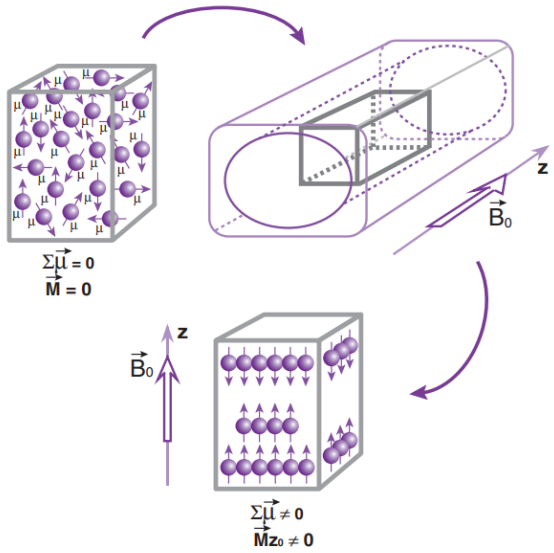


Figure 2.3: Illustration of the effects of applying a constant magnetic field $\overrightarrow{B_0}$ to the net magnetization vector $\overrightarrow{M}$. Source: [15]

B_0 and the temperature) is slightly above 1, meaning that there are slightly more parallel than anti-parallel protons. In our computations, we will only consider this surplus of parallel protons since the magnetic moments of all other protons cancel each other out. Moreover, their precession follows a particular angular frequency:

$$\omega_0 = \gamma B_0$$

(Larmor's equation) where ω_0 is the Larmor frequency proportional to B_0 and γ is the gyromagnetic factor (constant) specific to the isotope (e.g. $\gamma_{(^1H)} = 42.58$ MHz/T). At equilibrium, the macroscopic magnetization vector $\overrightarrow{M}$ is aligned on $\overrightarrow{B_0}$ (along the longitudinal z-axis): there is no component in the xy plane.

During an MRI session, this equilibrium is disturbed by the addition of energy coming from a spinning magnetic field (electromagnetic wave or radio frequency (RF) pulse) $\overrightarrow{B_1}$ in the xy plane. For this transfer of energy at equilibrium state to take place, the magnetic field's angular frequency ω_r must be equal to the Larmor frequency ω_0: this process is called resonance. In particular, when this criterium is met, $\overrightarrow{M}$ will also start precessing about $\overrightarrow{B_1}$ at the angular frequency of $\omega_1 = \gamma B_1$ additionally to the precession around the z axis. The RF pulse causes the proton to precess not only about Oz anymore but also about $\overrightarrow{B_1}$, its path following a spiral shape (Figure 2.4).

Depending on the RF pulse, it can induce the rotation of the magnetization vector to a certain angle. An RF pulse of $90°(\pi/2)$ has the following implications (Figure 2.5):

- Before the pulse, $\overrightarrow{M}$ is aligned on Oz so $\overrightarrow{M} = \overrightarrow{M_{z_0}}$ and $\overrightarrow{M_{xy}} = 0$

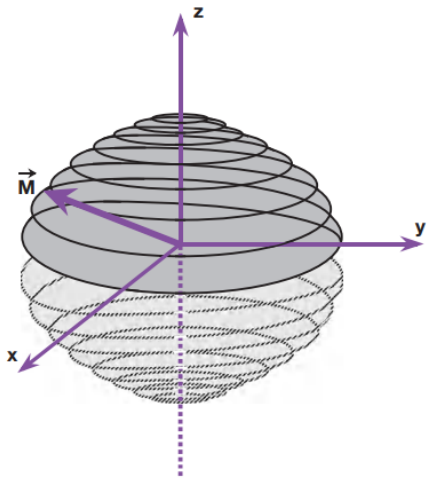


Figure 2.4: Spiral movement of two precessions described by the extremity of the magnetization vector $\overrightarrow{M}$. Source: [15]

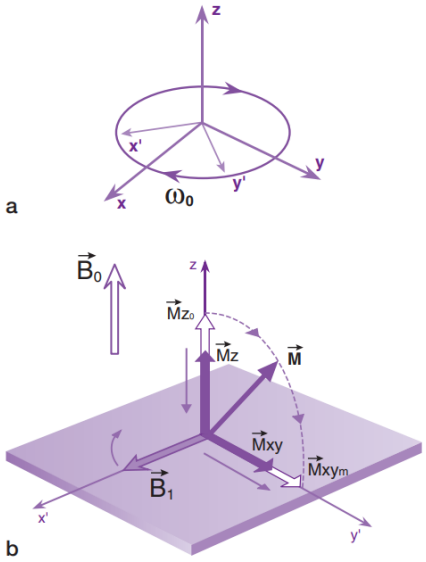
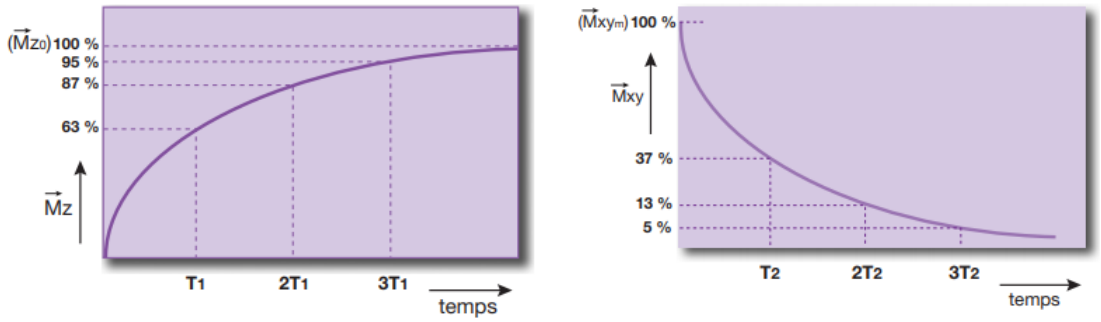


Figure 2.5: Illustration of the consequence of a 90°RF pulse on the magnetization vector $\overrightarrow{M}$. Source: [15]

- The RF pulse projects $\overrightarrow{M}$ in the xOy plane, which causes $\overrightarrow{M}$ to start precessing around $\overrightarrow{B_1}$. $\overrightarrow{M}z$ decreases and $\overrightarrow{M_{xy}}$ increases.
- At the end, $\overrightarrow{M}$ is completely in the xOy plane and its amplitude is equal to $\overrightarrow{M_{z_0}}$ (thus $\overrightarrow{M_{z_0}} = 0$).

2.1.3 The relaxation phenomenons

When the RF pulse stops, the net magnetization vector $\overrightarrow{M}$ follows the spiral path illustrated by Figure 2.4 to return to its original state of equilibrium (spinning



(a) (b)

Figure 2.6: Exponential curves of the net magnetization vector's longitudinal component growth w.r.t. T1 (a) and transverse component w.r.t. T2 (b). Source: [15]

around Oz). This step, called **relaxation**, is made up of two parts: the *longitudinal* relaxation, measured by T1 and the *transverse* relaxation, measured by T2.

- The longitudinal relaxation corresponds to the time it takes for the magnetization vector $\overrightarrow{M}$ to return to its original state (directed along Oz) after the RF pulse. It is thus also equivalent to the time it takes for $\overrightarrow{M}$ to lose its xy component. This relaxation corresponds to an exponential growth according to a constant time T1 (Figure 2.6a), specific for every tissue and corresponding to the time it takes for $\overrightarrow{M}$ to reach 63% of its final value.
- The transverse relaxation corresponds to the phase shift of the spins, causing the transverse (Oxy) component of the net magnetization $\overrightarrow{M}$ to disappear at an exponential rate (Figure 2.6b). This rate is also defined according to a constant T2 specific for every tissue. T2 is directly linked to inhomogeneities at a molecular level (small local magnetic fields superimposing themselves to $\overrightarrow{B_0} \implies$ phase shift of the spins).

T2* notion

To this point, we have always considered the magnetic field $\overrightarrow{B_0}$ induced by the magnet to be homogeneous. While this is true at a macroscopic scale, the story is a bit different at a microscopic scale. These inhomogeneities of "instrumental" origin will cause the phase shift to go faster. We use T2* to represent the conjunction of the effects of inhomogeneities at a molecular level (T2) and the ones induced by the instrument. Figure 2.7 illustrates this notion. The spin-echo sequence aims to get rid of these (constant) "instrumental" inhomogeneities due to the external magnetic field $\overrightarrow{B_0}$.

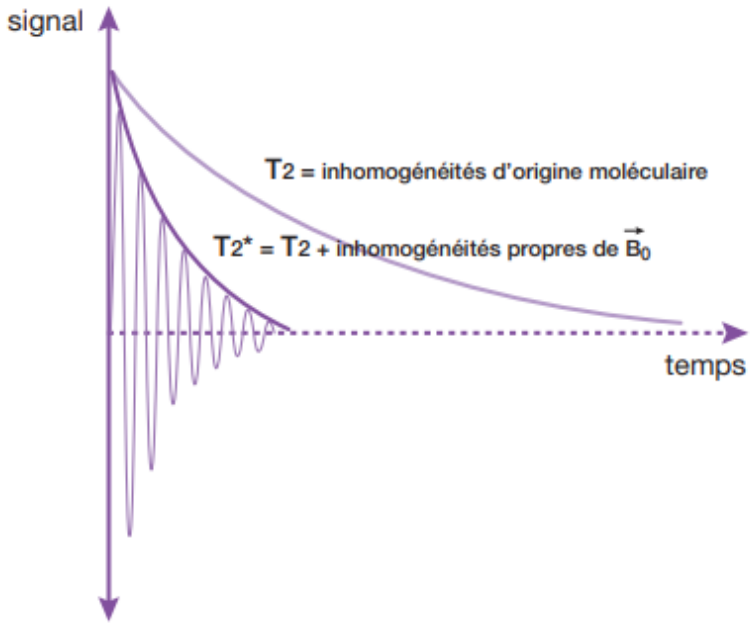


Figure 2.7: T2* notion. Source: [15]

2.1.4 The spin-echo sequence

The spin-echo sequence aims at discriminating the the two types of inhomogeneities of T2* to extract only T2. It is one of the first sequences developed (1955), well before the invention of MRI. The sequence follows these steps (Figure 2.8):

1. At time $t = 0$, we apply a 90 °pulse that throws the longitudinal component of the magnetization vector $\overrightarrow{M_{L0}}$ in the transversal planes. All spins are perfectly aligned and in phase, and the transversal magnetization is maximal ($\overrightarrow{M_T m} = \overrightarrow{M_{L0}}$).

2. The spins quickly dephase (due to the inhomogeineities of intrumental origin).

3. At time $t = TE/2$ (TE = Echo Time), we apply a 180°RF pulse

4. The spins rephase by spinning in the opposite rotational motion[1].

5. At time $t = TE$, the spins are rephased: the signal reappears in the form of an echo and can be measured. The spins are not completely in phase, but this phase shift is now entirely due to molecular inhomogeneities (and not instrumental ones).

This cycle allows to capture only one line of the image matrix (corresponding to the Fourier plane). To capture the other lines (a complete cycle is equivalent to a 128 or 256 lines image), one has to repeat it for each and every one of them. In this sequence of events, the echo time TE corresponds to the measure time, and the repetition time TR corresponds to the time interval between two pulses of 90°.

[1]Protons previously spinning clockwise now spin anti-clockwise and vice versa.
[2]Source: https://www.youtube.com/watch?v=WvOwHeaXZ-Y (visited on 24th May 2021).

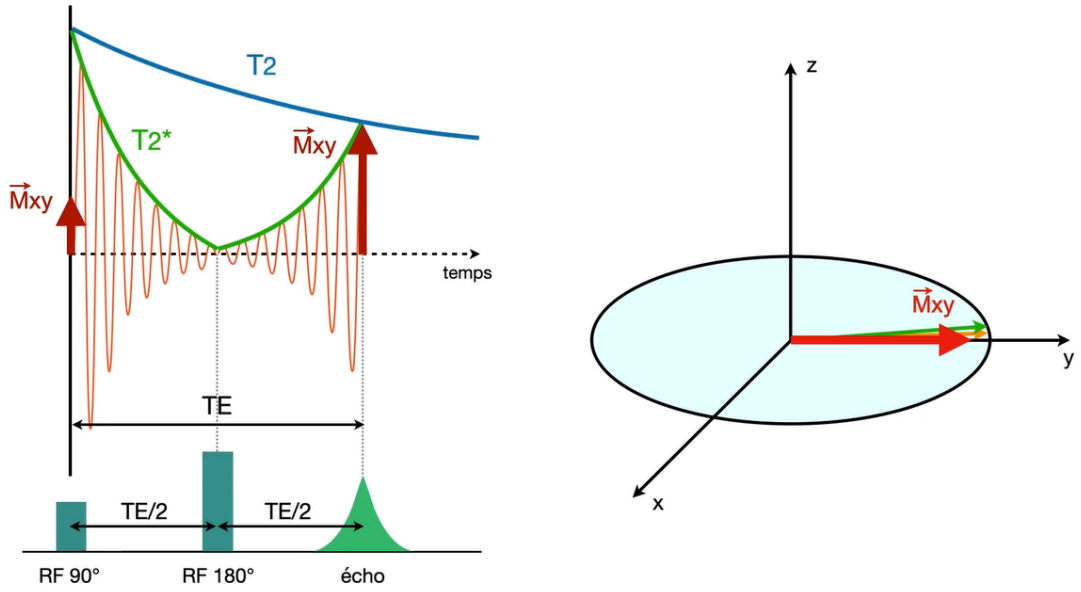


Figure 2.8: Chronology of events during the spin-echo phenomenon.[2]

2.2 Medical Context

2.2.1 Multiple Sclerosis

Multiple sclerosis [16] (MS) is a chronic immune-mediated disease of the Central Nervous System (CNS, includes the brain and the spinal cord). It affects more than 2 million people worldwide [17], and is the most common cause of non-traumatic disability in young adults. To this day, no cure has been found to treat it. Patients with multiple sclerosis experience episodes of neurological symptoms that are fully or partially reversible and which can last from a few days to a few weeks. Symptoms are diverse and typically include unilateral visual loss, limb paralysis, ataxia[3], weakness, sensory loss, or double vision [18]. These symptoms are generally related to the location of the lesions. After 10 to 20 years, the development of a progressive clinical course is typically observed in patients with MS, who ultimately end up with motor or cognitive disabilities.

The main pathologic mechanisms causing the clinical manifestations are inflammatory demyelination, the loss of the myelin sheath that surrounds nerve fibers mediated by inflammatory cells and axonal degeneration [19][20]. MS is subdivided into four clinical subtypes according to the progression of the disease [21] (see Figure 2.9):

- **Clinically isolated syndrome (CIS)**, representing the first clinical episode that may evolve or not to MS.
- **Relapsing-remitting MS (RRMS)**, characterized by clearly defined relapses (attacks) with full or incomplete recovery and with little to no disease

[3] Ataxia describes a lack of muscle control or coordination of voluntary movements

progression between attacks. This type accounts for the vast majority (85 to 90%) of MS cases at onset [22]. However, patients with RRMS will often evolve into a secondary progressive phase.

- **Secondary progressive MS (SPMS)**, concerns patients who first presented a relapsing-remitting form and in whom the relapses disappear after a certain time to give way to a slow, more or less regular progression of the disease.
- **Primary progressive MS (PPMS)**, characterized by a progression from onset where the evolution is quite slow, typically noticeable on the scale of years.

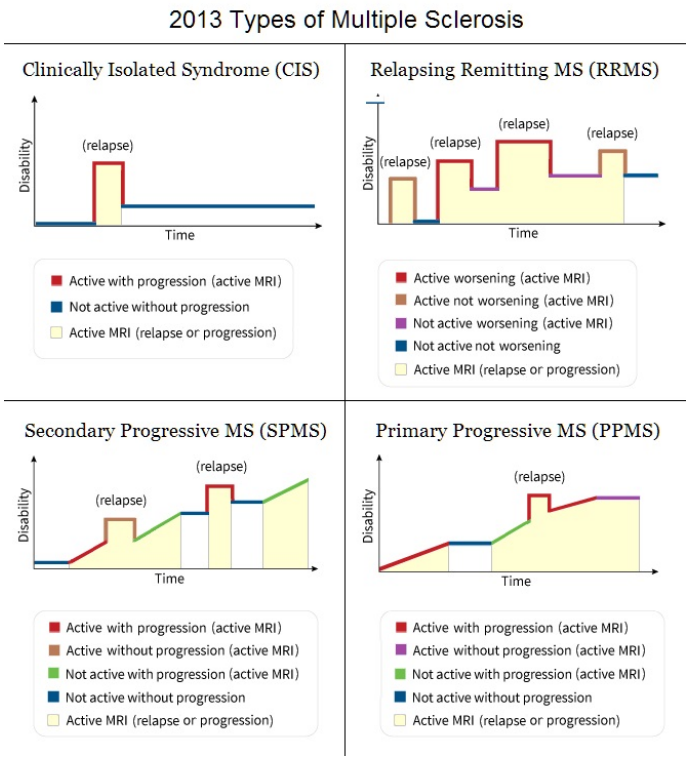


Figure 2.9: Subtypes of multiple sclerosis illustrated (2013). [4]

Pathological studies [16] provide evidence to suggest that MS is a heterogeneous disease, grouping together several different neurological diseases that share the same biological, clinical and imaging characteristics, forming the so-called "MS spectrum". In particular, both the disease course and the response to treatment in MS show heterogeneity at a patient-level. [23]

2.2.2 Diagnosis: The McDonald Criteria

Diagnosing MS requires the demonstration of disease dissemination in space and time using MRI, additionally to clinical and laboratory criteria (mainly the presence of CSF-specific oligoclonal bands - OCB). [24] On MRI, the spread in time can be demonstrated by the appearance of new hyperintense lesions on T2w or of a

contrast-enhancing lesion on post-gadolinium T1w images. The dissemination in space depends on the presence of at least one lesion in at least two territories typical of MS (periventricular, infratentorial, cortical/juxta-cortical and spinal cord). These typical territories are presented schematically on Figure 2.10. Additionally, MRI being the only method available to detect such lesions, it has recently become a major diagnostic tool in MS allowing an earlier, more sensitive, and more specific diagnosis supplementing clinical findings. Details on MR imaging in MS will be further discussed in section 2.2.4.

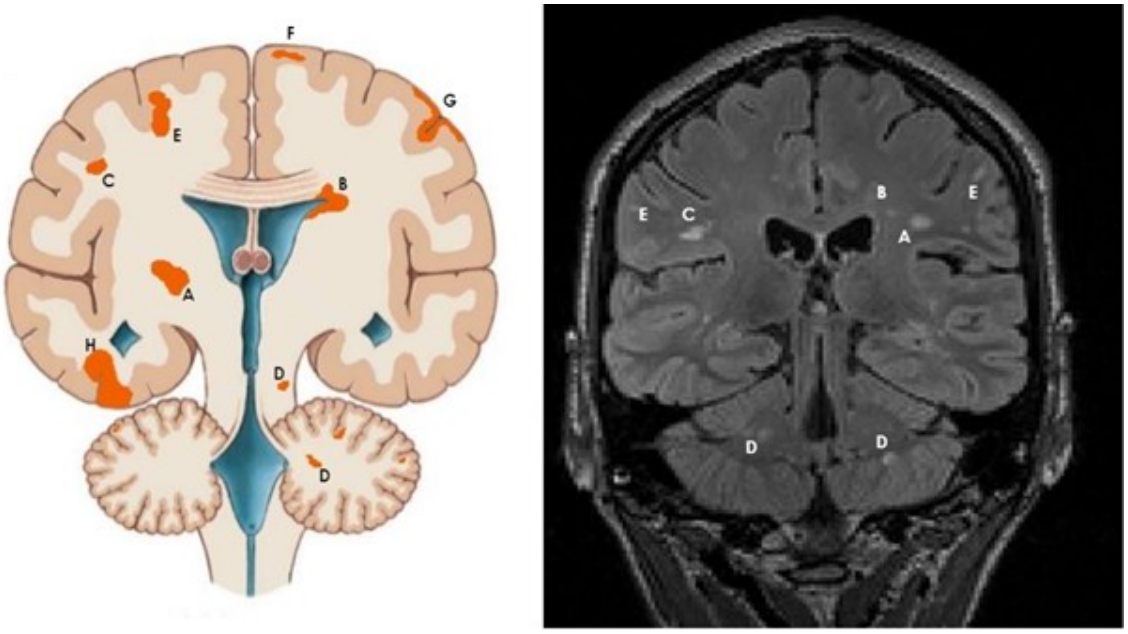


Figure 2.10: Schematic description of topography of multiple sclerosis lesions (left panel) and equivalent MRI of a coronal slice (right panel). Letters represent examples of different types of lesions (shown in orange in the scheme and as hyperintense signal in the MRI) according to their location: A - pure white-matter lesion; B - periventricular lesion; C - juxtacortical lesion; D - infratentorial lesions; E, F, G, and H - cortical lesions type I,II, III, and IV, respectively. The image on the left panel was adapted and reprinted by permission, of Pearson Education, Inc., New York, New York, from MARTINI, FREDERIC H.; TIMMONS, MICHAEL J.; TALLITSCH, ROBERT B., HUMAN ANATOMY, 7th, © 2012.

The McDonald criteria apply principally to patients with a typical CIS that could suggest the start of RRMS. However, they can also be applied on patients undergoing insidious neurological progression suggesting PPMS. Before stablishing a confident diagnosis of MS, one must ensure that there is no better explanation for the clinical symptoms first, and second that there is objective clinical evidence to confirm the presence of lesions in the CNS whose location corresponds to the attack. An important note is that the McDonald criteria, while performing well when MS is clinically suspected, are not intended for the differential diagnosis of MS

[4]Source: https://my-ms.org/ms_types.htm

(separate MS from other similar conditions, hereinafter referred to as "MS mimics"). A condensed view of the McDonald criteria can be found on Figure 2.11.

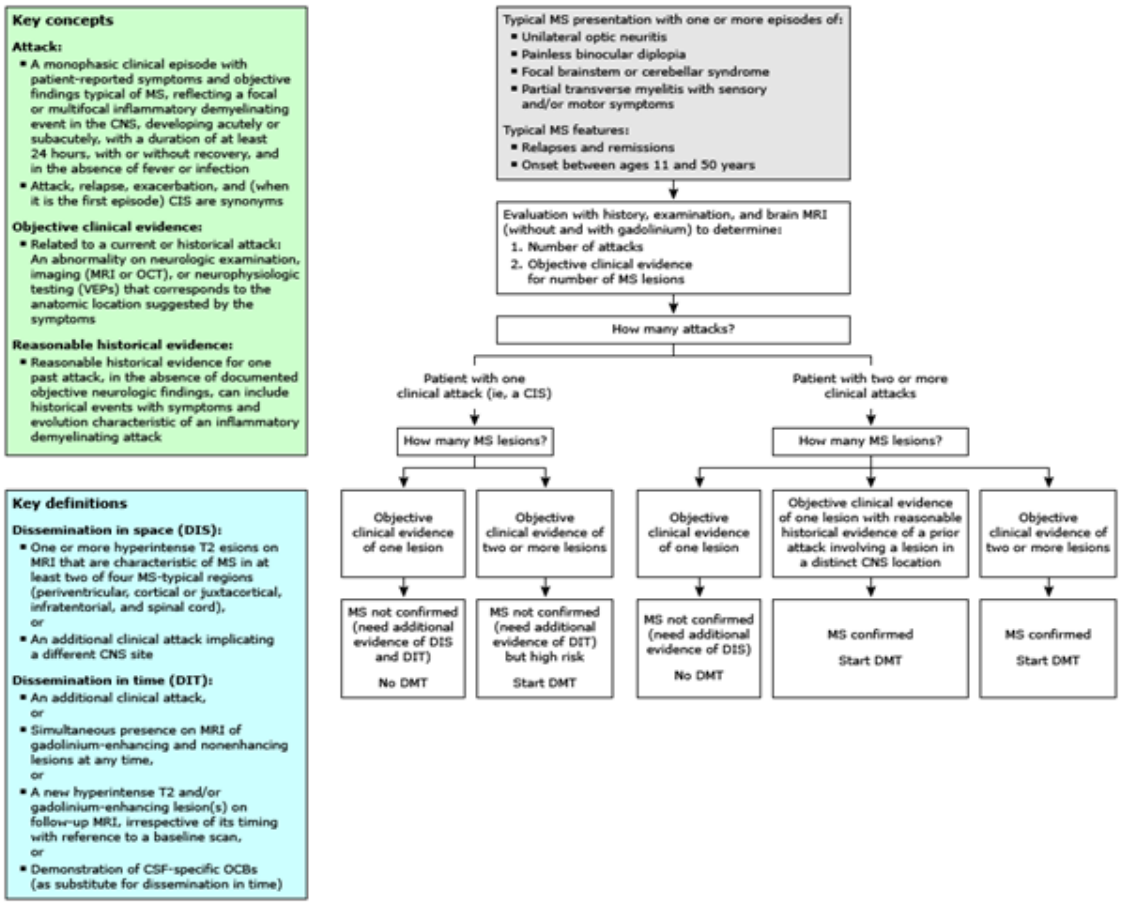


Figure 2.11: Condensed view of McDonald criteria for the diagnosis of MS. CIS: clinically isolated syndrome; CNS: central nervous system; CSF: cerebrospinal fluid; DIS: dissemination in space; DIT: dissemination in time; DMT: disease-modifying therapy; MRI: magnetic resonance imaging; MS: multiple sclerosis; OCBs: oligoclonal bands; OCT: optical coherence tomography; VEPs: visual evoked potentials.

However, and despite their utility, the above-mentioned MRI and diagnostic criteria (dissemination in timespace) are based on visual assessment of lesions using imaging protocols that have limited specificity for the underlying tissue properties. [2] Consequently, MS mimics can also meet these criteria [7], resulting in an inaccurate diagnosis and / or a potentially harmful treatment.[25] To meet this demand, biomarkers more specific to MS based on more advanced MRI features have recently been investigated. These biomarkers mainly include paramagnetic rim lesions [5] and the central vein sign. [2] Moreover, a very interesting and novel idea would be to use machine learning techniques such as classification and clustering on these advanced features combined with clinical and laboratory features to both better understand the differences between MS and its mimics, and potentially identify subclasses of the disease.

2.2.3 Disease Severity and Prognosis

Neurologic disability due to MS highly vary and can be described using several measures such as the frequency of relapses, the severity of symptoms and the rate of clinical worsening. In particular, a qualitative scale called the Expanded Disability Status Scale [26] (**EDSS**) has been established to assess patients' clinical disability in MS. This scale, illustrated on Figure 2.12, ranges from 0 for normal condition to 10 for death due to MS. Despite being the most widely used measure to assess disability, the EDSS is however not linear and is not able to distinguish patients who have accumulated disability quickly from those who have done so for decades. To meet this demand, another measure aiming to assess the *severity* of the disease emerged in 2005. [27] This score, called the Multiple Sclerosis Severity Score (**MSSS**), combines the EDSS with the disease duration and is very useful to classify MS patients based on their rate of progression.[28] Finally, yet another measure called Age Related Multiple Sclerosis Severity Score (**ARMSS**) aims to combine the age (typically unbiased and easily obtained) with the EDSS. ARMSS might be a more adaptable tool, reducing research biases and statistical power loss due to incorrect or missing onset dates. [29] One of the advantages of these last two measures is that, contrary to EDSS, they are *continuous* scores (and not scales), which can improve the the statistical power when detecting differences in disability between groups of patients.

The prognosis tries to predict the future course of the disease, and can be translated in MS in the prediction of a patient's EDSS after a certain time. While none are considered reliable by consensus and the outcome prediction for patients is rather limited, many indicators have been determined as potential prognostic factors estimating the chance of recovery from MS. Among them, there are MRI markers such as the lesion load (number of lesions) and cerebral atrophy (an exhaustive list can be found in section "MRI features for the diagnosis and prognosis of MS"). Classical MRI metrics include T1-w hypointense white-matter lesions (so called "T1 black holes") and spinal cord lesions, while more advanced metrics include the presence of cortical lesions, of paramagnetic rim lesions as well as quantitative measures of tissue damage. The lesions' locations have also been known to be an interesting prognostic factor. Figure 2.10 shows a schematic view of the topography of MS lesions.

2.2.4 Magnetic Resonance imaging in multiple sclerosis

To diagnose MS, several MRI sequences are available and used either in research or in the clinical routine. An MRI sequence is a particular setting of radiofrequency pulses and gradients resulting in images with a particular appearance. Among those used frequently in MS clinical practice or research, we find (not exhaustively):

- **T1-weighted sequences (T1w)** (short Echo Time (TE) and Repetition Time (TR)). Causes for hyperintensities ('brightnesses') on these images include but are not limited to the presence of fat, slow-flowing blood, calcium, copper, iron or paramagnetic contrast products (e.g. gadolinium). Therefore, MS lesions are typically hypointense (black) in T1w sequences and appear hyperintense after injection of gadolinium when acute inflammation is present.

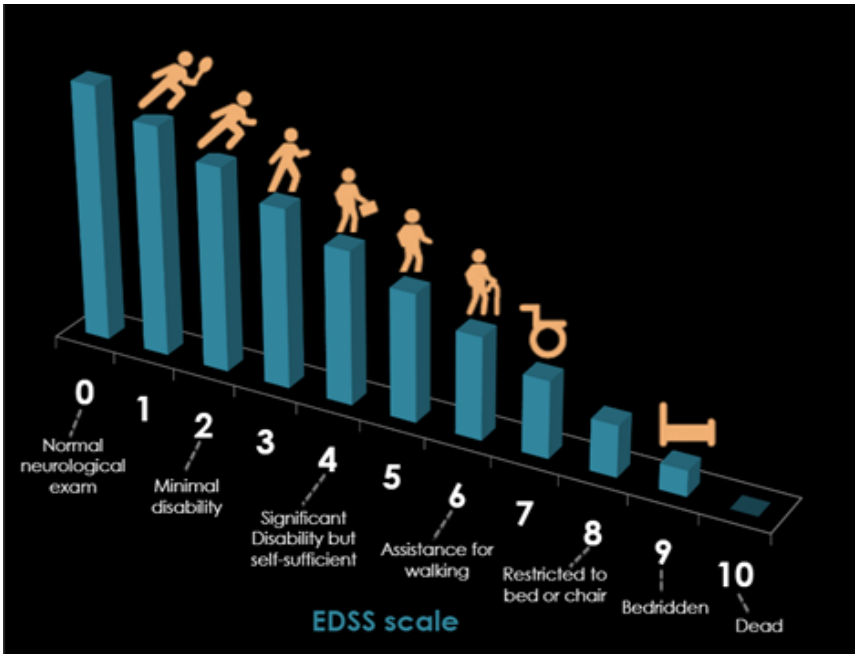


Figure 2.12: Schematic representation of the expanded disability status scale (EDSS). The scale ranges from 0 to 10, in 0.5 unit increments (small increments are not represented in this figure). From 1.0 to 4.5 MS patients are able to walk and the score attribution is determined on measures of impairment in several functional systems. From 5.0 to 9.5 patients are not able to walk, and in some cases restricted to wheelchair. Source : [30] .

- **T2-weighted sequences (T2w)** (long TR and TE, lower flip angle than T1w images). In these images, fluids (e.g. cerebrospinal fluid (CSF)) and fat appear white (hyperintense signal) while muscle tissues and grey matter will result in an intermediate signal intensity. White matter (WM) appears hypointense (darker) compared to gray matter (GM).

- **3D T2-FLAIR** (Fluid Attenuated Inversion Recovery), a T2w image that removes the signal from the CSF making it appear dark instead of bright. In particular, lesions appear bright [7] because they are mainly composed of water. Such technique is especially useful to detect hyperintense lesions around the ventricles (which appear bright/hyperintense on other T2w images).

- **3D-DIR** (Double Inversion Recovery), a T2w image where signals from CSF and WM are suppressed. These images are used to estimate the lesion load, to detect infratentorial or spinal cord lesions and, most importantly, to differentiate juxtacortical lesions from mixed WM-GM lesions and to detect purely intracortical lesions.

- **3D MPRAGE** (Magnetization Prepared Rapid Gradient Echo), is a sequence designed for rapid acquisition with T1w dominance. MPRAGE images are particularly useful and efficient for brain segmentation and volumetric analysis, using tools like Freesurfer's recon-all [31].

- **3D MP2RAGE**, a strongly T1-weighted image where the contrast between GM and WM is higher compared to MPRAGE. In particular, MP2RAGE images have been shown to be particularly useful in the detection of cortical lesions [32] and in giving quantitative estimates of T1 (and thus a possible measure of tissue damage).
- **Susceptibility weighted imaging**, an MRI technique taking advantage of the magnetic susceptibility differences between different compounds, creating new sources for MR contrasts. These images are obtained by Gradient Echo Imaging (GRE). In particular, this thesis works primarily with a sequence called *Echo Planar Imaging* (**3D EPI**), because it accelerates the sequence duration down to 4 to 5 minutes while achieving a sub-millimeter resolution in all 3 planes of the space. EPI is efficient and useful to study fine anatomical details in WM and GM, visualize parenchymal veins, and detect hemorrhage or calcification. [33] Additionally, we can obtain T2*-weighted (T2*w) and phase images from the EPI. Finally, we can also compute a quantitative susceptibility map (QSM) based on these images. As quantitative imaging methods are not very numerous in MS analysis, this map can give powerful results.

Regarding lesions and how to best assess them, Figure 2.13 offers a condensed view of the optimal sequences for each lesion type.

Table I Optimal imaging sequence suggested for each lesion type

Lesion category	Core sequence(s) for primary identification	Alternative confirmatory sequence(s)
Core lesional features for diagnostic criteria		
Periventricular	T_2-FLAIR (preferably 3D)	T_2-weighted, PD-weighted, 3D T_1-weighted MPRAGE
Juxtacortical/Cortical	T_2-FLAIR (preferably 3D) (Cortical: DIR)	3D T_1-weighted MPRAGE, T_2, DIR, PSIR (Cortical: 3D T_1-weighted MPRAGE, PSIR; T_2-FLAIR less optimal)
Infratentorial	T_2-FLAIR (preferably 3D) (Bink *et al.*, 2006; Gramsch *et al.*, 2015; Moraal *et al.*, 2008; Wang *et al.*, 2018)	T_2, PD, 3D T_1-weighted MPRAGE
Spinal cord (cervical + thoracic)	⩾2 sagittal sequences including STIR, T_2, PD, PSIR or 3D T_1-weighted MPRAGE	Axial T_2 (Weier *et al.*, 2012)
Gadolinium-enhancing lesions	Mildly/moderately T_1 SE or GE after a single dose gadolinium-based contrast agent with ⩾5-min delay —avoid heavily 3D inversion-prepared T_1-weighted MPRAGE —no MT pulse	Pre-contrast T_1 (optional)
Additional multiple sclerosis lesions not currently included in formal diagnostic criteria		
Optic nerve	2D STIR (coronal) Post-contrast fat-suppressed T_1 (axial and coronal)	2D FSE (coronal) 2D STIR (axial) Alternatives (good contrast but lower resolution): 3D DIR, 2D/3D FSE T_2, 2D/3D fat suppressed T_2-FLAIR
Future pathophysiology-based characteristics		
Central vein sign	3D T_2* (with segmented EPI) T_2-FLAIR* (T_2-FLAIR + T_2* with segmented EPI)	SWI
Subpial demyelination	7 T T_2* or MP2RAGE	PSIR and/or 3D T_1-weighted MPRAGE; T_2-FLAIR less optimal; DIR
Smoldering/slowly expanding lesions	Phase of 7 T T_2*-weighted GRE	Phase of 3 T 3D T_2* or SWI Longitudinal T_2 or T_1 images

DIR = double inversion recovery; EPI = echo-planar imaging; FSE = fast spin echo; GE = gradient echo; GRE = gradient recalled echo; MPRAGE = magnetization-prepared rapid gradient echo; MT = magnetization transfer; PD = proton density; PSIR = phase-sensitive inversion recovery; SE = spin echo; STIR = short-tau inversion recovery; SWI = susceptibility-weighted imaging; T_2-FLAIR = T_2-fluid-attenuated inversion recovery.

Figure 2.13: Optimal imaging sequence suggested for each lesion type in MS. Source: [7]

2.2.5 Advanced MRI features for the diagnosis and prognosis of MS

While several biological features exist for the diagnosis and prognosis of MS (like the presence of OCB in CSF, this work will mainly focus on the radiological features. Depending on the images acquired/available, one could consider the following MRI features:

- **Lesion** number, volume, shape, and location (3D FLAIR and MP2RAGE if available)
- Presence of the **Central Vein Sign** (CVS) assessed on T2*w images and which corresponds to the presence of a vein at the center of a brain WM lesion
- Presence of **Paramagnetic Rim Lesion** (PRL or Rim). PRLs feature a hypointense rim on phase images and a lesion parenchyma isointense to extralesional WM [34]
- **Brain morphometry** (3D MPRAGE with/without 3D FLAIR, or MP2RAGE if available). Brain structures mostly targeted for the volumetric analysis are WM, GM, thalamus, ventricles, caudate and putamen.
- Presence of **cortical lesions**, assessed on 3D-MP2RAGE and 3D-DIR sequences.
- **Tissue integrity**, assessed on T1 relaxometry maps obtained from the 3D-MP2RAGE,on Quantitative Susceptibility Mapping (QSM) obtained from the susceptibility-weighted-sequence or multi-shell diffusion MRI. [35]

The following sections will further explain the definition, methods of detection and relevance of some of the advanced MRI features.

2.2.5.1 Paramagnetic Rim Lesions

Definition

Studies have shown that more than half of MS lesions are chronic active.[36] Paramagnetic rim lesions (PRL or rims) can be visualized in vivo on susceptibility-based MRI as non-gadolinium enhancing lesions with a paramagnetic rim. These lesions, which are also known as “chronic active lesions/smoldering” [6] or “slowly expanding” [13] lesions, are defined by a slow rate of increase in size and a continuous tissue damage on MRI and are pathologically characterized by an inflammatory edge of activated microglia/macrophages containing iron.[37] [38] An illustrative example of such PRL can be found on Figure 2.14.

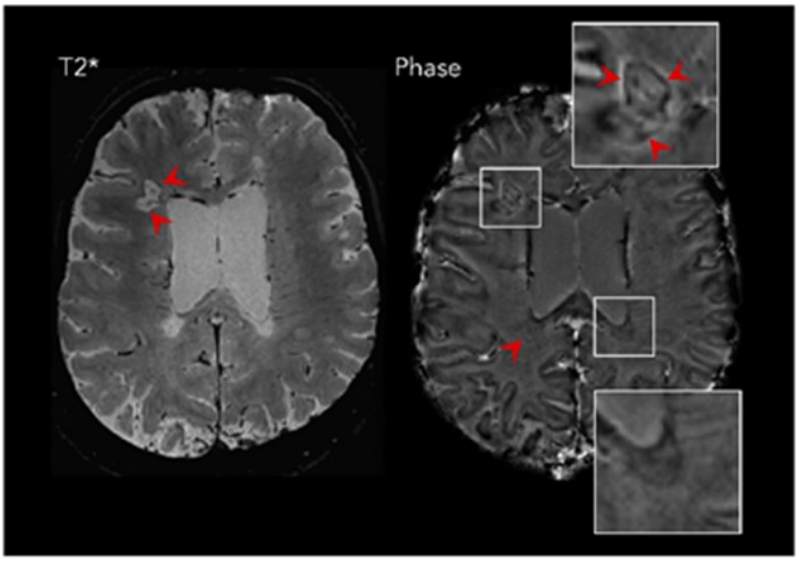


Figure 2.14: Representative MRI images of PRL in MS. Axial 3-tesla T2*-weighted magnitude and phase MRI images from a relapsing remitting MS patient (40 years old, male) showing PRL (arrowheads). Insets show magnified views. Adapted from [34]

.

Assessment

As previously stated, T2 and T1-weighted sequences (such as FLAIR and MPRAGE) are the main imaging techniques used for in vivo assessment of MS brain lesions, including lesion detection, count and volume estimation. Additionally, for PRL analysis, a susceptibility-based MRI sequence (like T2*-weighted magnitude and phase, susceptibility-weighted imaging, or quantitative susceptibility mapping) can be added to the imaging protocol, in order to classify lesions based on whether a paramagnetic rim is visible (PRL+) or not (PRL-).[39][40] So far, this PRL +/- classification is done through visual, operator-dependent inspection of the MRI susceptibility contrasts by experts.

Recently, however, a computer assisted method for PRL+/- classification called RimNet[10] has been proposed. This method is based on a 3D patch-based convolutional neural network (CNN) architecture, which exploits several MR imaging contrasts combined at the first and last layers of the network. Moreover, RimNet is to my knowledge the only deep learning method specifically developed to support PRL assessment in MS.

Relevance

Previous studies have shown that accrual of MS patient's disability [5][40][41] and in vivo neurodegeneration are associated with a higher PRL burden. [34]

2.2.5.2 Central Vein Sign

Definition

It has been known since 1820 that MS lesions develop around small parenchymal vessels (mainly veins and venules).[42] From a physiopathological point of view, the perivascular space around these veins is a prime site for immune cells to interact with antigen-presenting cells, which can consequently activate an inflammatory cascade that leads to the development of the lesion around the vein.[43][44][45] Today the presence of a vein at the center of MS lesions can be depicted in vivo using susceptibility weighted techniques. The so called "Central Vein Sign" (CVS) on MRI represents the presence of a vein at the center of a brain WM lesion. Figure 2.15 shows an example of the CVS.

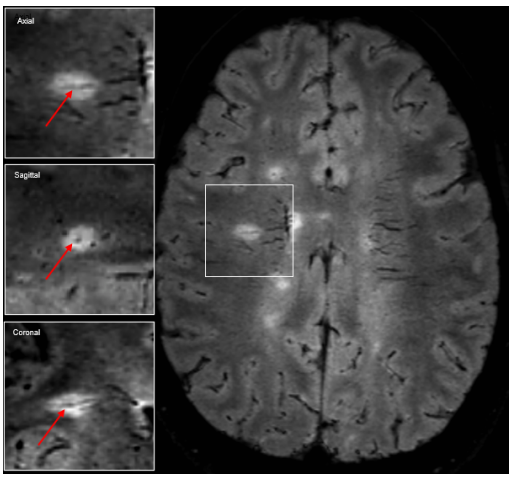


Figure 2.15: Representative MRI images of the CVS in MS. Axial 3-tesla FLAIR* showing CVS (arrowheads). Insets show magnified views. Adapted from [2].

Assessment

After lesion segmentation, the presence of the CVS can be assessed on T2*w images. As it is the case with PRLs, the classification of lesions with or without the CVS is also done through a visual inspection of the images by experts, and a deep-learning based prototype for the automatic assessment of the CVS called "CVSnet" has recently been developed.[9]

Relevance

Recent studies have shown that MS has a higher frequency of perivenular lesions than other MS mimicking diseases.[2] The CVS is associated to a high specificity and a moderate sensitivity in differentiating MS from its mimics [4]. Moreover, the CVS may be a useful tool to diagnose MS in patients with atypical clinical or radiological presentations [3] and to predict the development of MS in at risk individuals [46]. Finally, the amount and percentage of CVS NEGATIVE lesions in a patient brain

can give us an idea of the microangiopathic/ischemic brain lesion load, an innovative field of research in MS. [14][47]

2.2.5.3 Cortical Lesions

Definition

Cortical lesions are defined as focal lesions either directly adjacent to the cortex (*juxtacortical* lesions), involving both the cortex and the normal WM (*leukocortical* lesions) or located completely in the cortex (*intracortical* lesions). [7] All cortical lesions must meet two criteria (on DIR images): first, correspond to a hyperintense area (compared to the surrounding GM) on a T2w image and second be at least 3mm long on the main in-plane axis. [48]

Other lesions caused by MS mimicking disorders (such as migraine or Neuromyelitis optica) can also be located close to the cortex. These lesions are however typically situated in the deep WM and feature the presence of normal WM between the cortex and the lesion and are thus called "sub-cortical lesions"

Assessment

The detection of cortical lesions on MR images is quite challenging. T2-FLAIR images (preferably 3D) are commonly used to detect them in the clinical routine, although other -more sensitive- sequences can be used. Among those, DIR sequences are frequently used, although one can also visualize them using PSIR, MPRAGE or MP2RAGE sequences. Interestingly, techniques were recently developed to automatically detect cortical lesions based on the combination (or not) of DIR with other sequences like MPRAGE, T2-FLAIR or MP2RAGE.[49][50]

Relevance

Although only juxta-cortical lesions (and not purely intracortical) can be easily depicted using clinical-routine MRI sequences (like FLAIR), the cortical location is considered typical of MS and was recently included in the MS diagnostic criteria. [24] (see Figures 2.11 and 2.10). Cortical lesions are also an important tool for the differential diagnosis of MS. Indeed, these lesions have a quite high specificity for MS which makes them particularly useful for the differential diagnosis. Additionally, cortical lesions can improve the prediction of conversion from CIS to MS. [12][51][52] [7] Finally, other studies have shown that GM damage is associated with physical and cognitive disability progression in MS, and thus a higher EDSS. [53]

Chapter 3

Technical tools and methods used in this work

A large number of image processing and data analysis tools were adopted for this thesis, some of them (like image registration or reorientation) were almost used on a daily basis, while others were used in a more sporadic way. This chapter aims at showcasing the most relevant.

3.1 Image processing

After image acquisition, a series of DICOM[1] files are obtained. According to the official site of the DICOM standard [54], this type of file is the international standard for medical images and related information. Starting from the DICOM format, neuroimaging researchers usually transform the data in order to obtain a more suitable format for image processing and storage. Since for this thesis, this transformation had to be done for the whole dataset (about 150 scans), we aimed at automating the overall processes, as well as optimising data organization. Hereafter you will the tools, methods and standards or good practices that helped me a lot during the image processing part of my thesis.

3.1.1 DICOM to NIfTI converter

Although the NIfTI[2] file format was originally created for functional neuroimaging data storage and display, it is also actually quite useful in our case. Indeed, while the DICOM format is made of a lot of files in a directory, the NIfTI format wraps up all the useful information into a unique file which can be additionally compressed (making it a compressed NIfTI file). NIFTI files are simpler and easier to manage (as compared to DICOM files) and represent a substantial saving of memory space. NIfTI files extensions are .nii and compressed NIfTI files end with .nii.gz .

I chose to use one of the most widely used converters, *dcm2niix* [55] to accomplish this task. Additionally to 'simply' converting data, this tool also very conveniently

[1]DICOM stands for Digital Imaging and Communications in Medicine
[2]NIfTI stands for Neuroimaging Informatics Technology Initiative

anonymizes the data and produces a *json* file including a subset of the DICOM's metadata. Moreover, it can also be easily integrated into a pipeline as it can run in a bash shell or with Python.

3.1.2 Brain Imaging Data Structure

Brain Imaging Data Structure (BIDS) aims at being an internationnal inter- and intra-lab consensus for file naming and structuring in brain imaging. It aims at avoiding misunderstandings and wasted time on rearranging data or editing scripts expecting a certain kind of structure.[56] In particular, all images in the BIDS format must be NIfTI's, accompanied by a *.json* file with all metadata that could not be contained in the NIfTI header. As an illustration of a database in BIDS format, Figure 3.1 shows a diagram of the structure used during this thesis. Note that all type of images and files have a particular naming pattern.

One may refer to the official BIDS documentation for additional details.[57]

3.1.3 Registration

Image registration is defined as a procedure which superimposes two or more images from, for example, two different sources, or from the same scene to geometrically align the images to be analyzed.[58] In our case, the images are either taken at different times (which could imply the patient has moved between the two image acquisitions, even if it happens in the same acquisition session) or are just by nature different, like the images extracted thanks to two different MRI sequences.

To illustrate this concept, Figure 3.2 shows an example of the registration of two images of a brain. At the end, every voxel[3] in both images should correspond to the same coordinate point in space.

While several registration tools exists in neuroimaging (to cite only a few: FSL's Flirt, Elastix, AIR, ...), I chose to use the ANTs registration tool [60]. This choice was mainly supported by [61] in which the authors compared the performances of five registration tools and concluded that ANTs gave the best accuracy. Another important aspect that has been taken into account in the choice was the fact that it runs with bash commands, which makes it very easy to include in an automated pipeline. One drawback however of using this tool is that even if a work exists to integrate it as a Python module, it can at the moment only be run on a Linux system. However, as many other important neuroimaging tools such as Freesurfer [31] also only run on Linux systems, it does not pose much problem.

Different types of registration exist, each one with their own purpose. To cite only a few, there is the translation transformation, the "similarity" transformation, consisting of scaling, rotation and translation, and the rigid transformation, consisting of rotation and translation only (an exhaustive list of what ANTs offers can be found

[3] A voxel is a 3D pixel

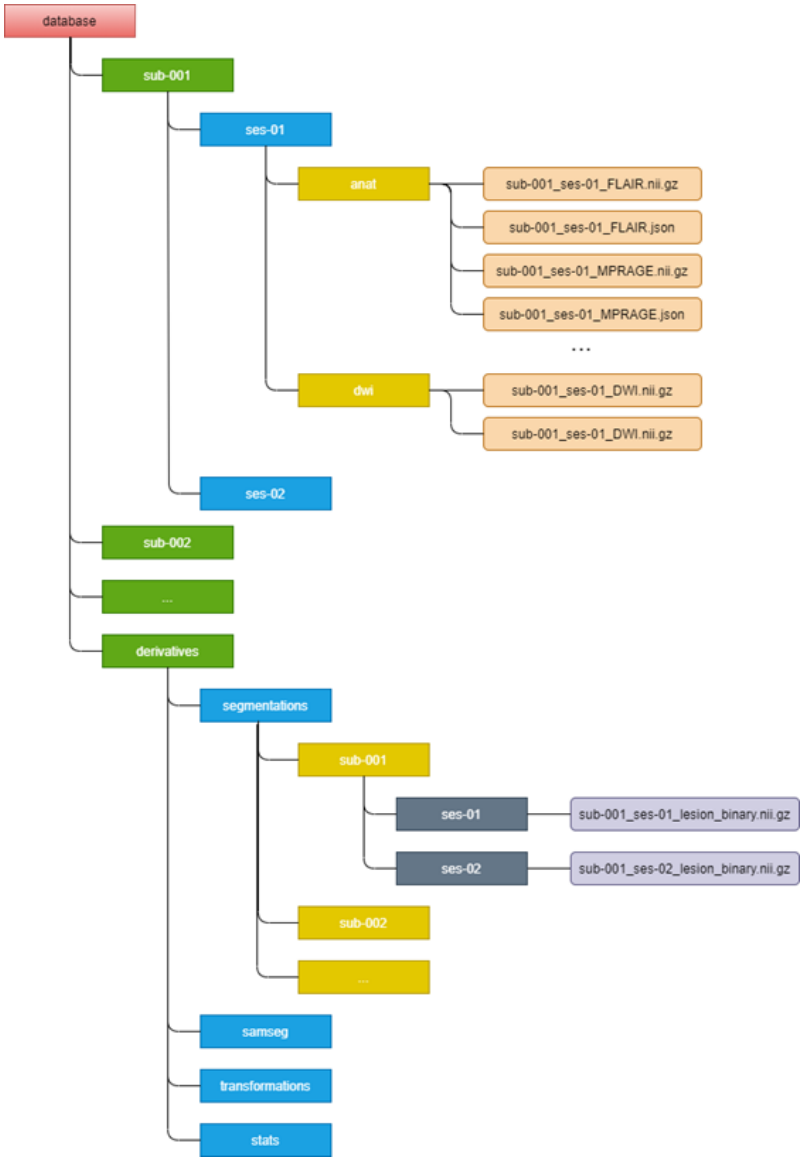


Figure 3.1: Diagram of BIDS structure

in their documentation[4]). In our case, only the latter was needed and used, since the images to be registered were consistently acquired on the same patient.

3.1.4 Segmentation

According to [62], image segmentation is one of the most crucial and practical tasks in medical imaging. It aims at classifying the pixels (voxels in our case) into several labels, each one corresponding to a different entity (i.e. anatomical). In neuroimaging, it is mainly used to calculate the volumes of the different anatomical structures of the brain, but also to visualize them, analyze their changes over time or to delimit pathological regions (lesions in our case). An illustrative example of brain tissues segmentation is shown on Figure 3.3.

[4]https://github.com/stnava/ANTsDoc

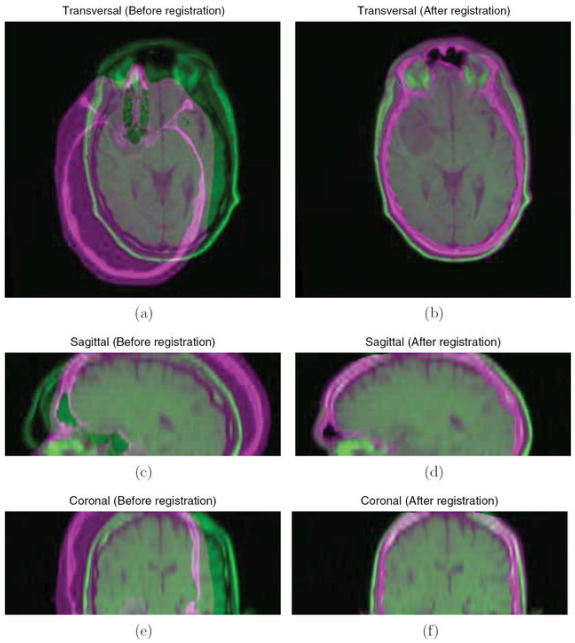


Figure 3.2: Example of registration of the brain. Adapted from [59]

(a) (b)

Figure 3.3: MPRAGE image (a) and whole-brain and lesion segmentation (b) of a patient diagnosed with multiple sclerosis.

Because image segmentation is a wide subject of research within itself and I did not have the time to implement it myself, I chose to use already validated methods during the course of my thesis. The most widely used and validated program performing brain parcellation is the main feature of FreeSurfer [31] called "recon-all". However, while this program is arguably the most widely used for this task, I chose to use a newer (available since 2020, published in 2021) segmentation method called SAMSEG[5] specifically designed for whole-brain and lesion segmentation in MS patients[63]. The advantages of using SAMSEG compared to FreeSurfer's recon-all were: 1) SAMSEG was specifically developed to segment MS brains (which is not the case of Freesurfer), 2) SAMSEG performs a very accurate WM lesion segmentation, which is also used/corected-for during the whole-brain volume computation (contrary

[5]SAMSEG stands for Sequence Adaptive Multimodal SEGmentation

to FreeSurfer), and 3) SAMSEG takes significantly less time than FreeSurfer (30min-1h30 vs 6h-24h). Moreover, a recent paper published in the prestigious journal *Neuroimage* showed that SAMSEG's brain segmentation is accurate and can be put to the same level as FreeSurfer.[64] Finally, it is important to note that SAMSEG, being integrated in the FreeSurfer framework, facilitates a lot of operations, works with bash commands and can be easily added to a pipeline.

SAMSEG outputs a set of files:

- *seg.mgz*[6] : the 3D segmentation where each voxel corresponds to an anatomical structure. This file is the equivalent of the *aseg.mgz* file output by FreeSurfer's recon-all.
- *samseg.stats*: This file comprises the estimated volume in mm^3 for each segmented structure.
- A set of files including volumes of the posterior probability for each brain structure.

3.1.5 Miscellaneous

Among the other tools that were used in the pre-processing, I also wanted to cite a few particularly useful Python packages: *pydicom*[65] (reads and handles DICOM files and file structures) *nibabel*[66] (reads NIfTI images and allows the conversion of the image into a numpy matrix) and *scipy*[67] (very useful for basic image processing like labelling). These packages allow to easily correct defect in the original images. This is very important because defects like the orientation are a frequent source of registration and/or analysis errors and can generate a huge waste of time if not detected and fixed.

Another tool I used is FreeSurfer's motion correction [68]. This tool corrects small motions between two or more images and outputs a standardized 255 cubed image with 1mm isotropic voxels. That standardized output is also reduced in terms of memory space, which was particularly useful for running SAMSEG, as SAMSEG alone requires a lot of RAM.

3.2 Data processing and analysis

Even though MRI features extracted through image processing represent the majority of the information included in this thesis' database, many other demographic, clinical and biological features must be extracted in order to fill-up the database. These clinical data came from many different sources and most of them had to be processed by a computer to avoid human encoding error and to save time. Moreover, the whole database had to be processed, analyzed and visualized before applying machine learning techniques. Hereafter you will find a list of useful Python packages I used to these purposes.

[6].mgz files are compressed .mgh files, which are used to store high-resolution structural data and other data which are to be overlaid on the high-resolution structural volume.

- *pandas* : pandas is a "fast, powerful, flexible and easy to use open source data analysis and manipulation tool". [69]
- *numpy* : numpy provides many tools for image processing, matrix computations, signal processing, statistical computating, and is written in compiled C which makes it very fast. [70]
- *sklearn* : scikit-learn regroups a set of simple and efficient tools for machine learning and data analysis. [71]
- *scipy* : scipy is an library for mathematics, science and engineering. [72]
- *seaborn* : seaborn is a Python data visualization library based on matplotlib. It has a high-level interface for creating visually appealing and instructive statistics visuals. [73]

3.3 Exploratory Data Analysis

Data scientists use exploratory data analysis (EDA) to examine and study data sets and describe their major features, frequently using data visualization techniques. It assists data scientists in determining how to effectively manipulate data sources to obtain the answers they want, making it simpler for them to find patterns, test hypotheses, and verify assumptions. EDA is generally used to examine what data might reveal outside of formal modeling or hypothesis testing tasks, and to gain a better knowledge of data set variables and their interactions. It might also assist you in determining whether the statistical approaches you're contemplating for data analysis are suitable.

3.3.1 Normality tests

The test of normality is a crucial step in determining the measures of central tendency and statistical methods for data analysis for continuous data. In the case where our data follows a normal distribution, parametric tests must be employed to compare the groups; otherwise, nonparametric approaches must be used. While there are visual methods to assess the normality of a distribution, such as Q-Q plot or histograms, I chose to focus on two specific numerical methods because it was easier to implement in an automatic pipeline. If one of these two tests failed, the distribution was characterized as *not* following a normal distribution.

- **D'agostino's k-squared test:** This method tests the null hypothesis that a sample was drawn from a normal distribution by analyzing the distribution's kurtosis and skewness. [74] [75]
- **Shapiro-Wilk test:** This test computes a statistic W:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i are the ordered random sample values and a_i are constants generated from the covariances, variances and means of the sample (of size n) from a normally distributed sample. [76]

Both of these tests provide a statistic associated to a p-value. If the p-value is inferior to an α-level of 0.05, the test fails to reject the null hypothesis that the sample was drawn from a normal distribution and the distribution is considered to be normal according to the test. Otherwise, the distribution is considered to follow another distribution than a Gaussian distribution. Moreover, it is important to note that both of these tests present a a bias associated to the sample size.

3.3.2 Bivariate analysis

One of the simplest types of quantitative (statistical) analysis is bivariate analysis. It entails the examination of two variables in order to determine their empirical relationship. Because a variable can either be quantitative or qualitative (categorical), there are three cases to handle when investigating the relationship between two variables.

3.3.2.1 Linear relation between quantitative variables and other quantitative variables

In order to measure the strength of a linear relationship between two quantitative variables, the main method is to analyze their correlation. The (Pearson) correlation is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where *cov* is the covariance and σ_A is the standard deviation of the quantitative variable A. A *positive* correlation implies that when one variable increases, the other will tend to increase as well, while a *negative* correlation implies that the other will decrease. In particular, a correlation of 1 indicates that there is a perfect positive correlation between two variables, -1 a perfect negative correlation and 0 no correlation at all.
Additionally, we may also take a look at the coefficient of determination or R^2 (equal to the square of the correlation coefficient) to understand which proportion of the variance can be explained from the independent variable(s). A heatmap of all R^2 can then be drawn to have a quick and easy look at the (linear) relationships between quantitative variables. Finally, to better visualize these relationships with a human eye, we can draw a scatter plot and try to fit a line to the data points (simple linear regression with ordinary least squares).

3.3.2.2 Relation between quantitative variables and categorical variables

Different methods exist to test whether a the different categories of a qualitative variable have an effect on a quantitative variable, each one associated to a particular case. Figure 3.4 shows which test to choose for each one of them. The first question we want to ask ourselves is how many categories or values our qualitative variable can

take: either 2 or strictly more than 2. For the next step, we have to choose between a parametric test, making assumptions about the parameters of the distribution data from which the sample is drawn, or a non-parametric test for which no assumption is made about the distribution of the data. In order to do this, we can use the normality tests described in section 3.3.1 on all subdivisions of the data according to the categorical variable's value (e.g. separating the weights drawn from the male population from the ones drawn from the female population and perform the tests on each of the two sub-samples when studying the effect of gender on a person's weight). If all subdivisions pass the normality tests, we can use the Student's t-test (2 samples) or an ANOVA (≥ 3 samples), otherwise we use the Mann-Whitney U test (2 samples) or the Kruskal-Wallis test (≥ 3 samples). Note that there is also a third step consisting of knowing whether our data is paired or not, but in our case the data we want to analyze is never paired.

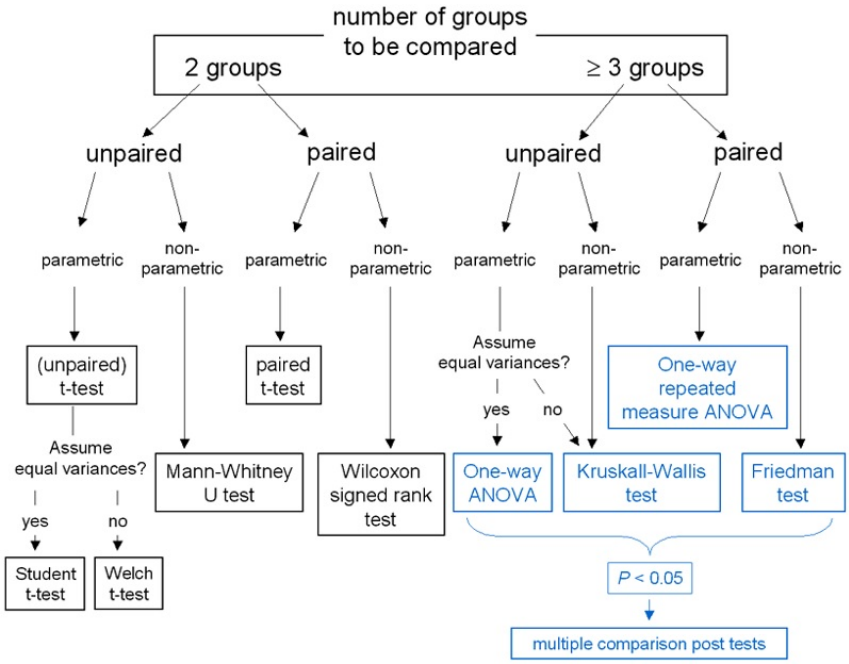


Figure 3.4: Flowchart showing what test to use when comparing distributions drawn from different populations[7]

The results of these tests may be accompanied by an illustrative boxplot.

3.3.2.3 Relation between categorical variables and other categorical variables

To investigate whether two categorical variables are independent or not, we typically use a χ^2-test (Chi-Squared). This test is based on the contingency table of the two variables, which corresponds to a frequency table of each pair of values the two variables can take. These contingency tables can be visualized with a heatmap with the associated result of the χ^2-test.

[7]Source: https://www.pinterest.es/pin/638526053388897645/

3.3.2.4 Non-linear relations between variables

As an alternative to the linear methods such as the correlation, we may use another measure called the *Mutual Information* to capture non linear relationships between two variables. The mutual information measures the "amount of information" gained about one random variable when knowing the other random variable.

Chapter 4

Data Acquisition and Pre-processing

4.1 Population

A total of 135 patients diagnosed with MS were included in this project, originating from two datasets. Images from Dataset I originated from 59 patients (35 female/ 24 male, mean age 47 ± 11 years, range [26 - 77]) treated in Lausanne, Switzerland. According to the McDonald's diagnosis criteria [24], 39 were diagnosed with RRMS, 10 with PPMS and 10 with SPMS. The mean EDSS score was 3.2 ± 2.2, ranging from 0 to 7.5. Images from this dataset were acquired on two 3T MRI scanners manufactured by Siemens (MAGNETOM Prisma and MAGNETOM Skyra, Siemens Healthcare, Erlangen, Germany). The following sequences were acquired: 3D-EPI (resolution = 0.65 mm^3; TR,TE = 64, 35 ms; acquisition time = 8 min); 3D-FLAIR (resolution = 1 mm^3; TR,TE,TI = 5000, 393, 1800 ms; acquisition time = 8.5 min); 3D MPRAGE (resolution = 1 mm^3; TR,TE,TI = 2000, 2, 1100 ms; acquisition time = 8 min); classic T2 (TR,TE = 3380, 10 ms; acquisition time = 8.5 min).

Images from Dataset II originated from 76 patients (43 female/ 31 male (2 unknown), mean age 47 ± 12.5 years, range [20 - 75]) treated either at ULB - Erasmus Hospital (37) or at the UCLouvain - Cliniques Universitaires Saint-Luc (39), Brussels, Belgium. According to the McDonald's diagnosis criteria [24], 32 were diagnosed with RRMS, 15 with PPMS and 29 with SPMS. The mean EDSS score was 3.8 ± 2, ranging from 0 to 7. Imaging was performed on two different 3T MRI scanners of the same model (Philips Ingenia). The following sequences were acquired: 3D-EPI (resolution = 0.55 mm^3; TR,TE = 63, 29 ms; acquisition time = 10.5 min); 3D-FLAIR (resolution = 1 mm^3; TR,TE,TI = 4800, 326, 1600 ms; acquisition time = 10 min); 3D MPRAGE (resolution = 1 mm^3; TR,TE = 8, 3 ms; acquisition time = 10.5 min); classic T2 (TR,TE = 4000, 80 ms; acquisition time = 10 min); T1w post-gadolinium (TR,TE = 324, 4 ms; acquisition time = 10 min).

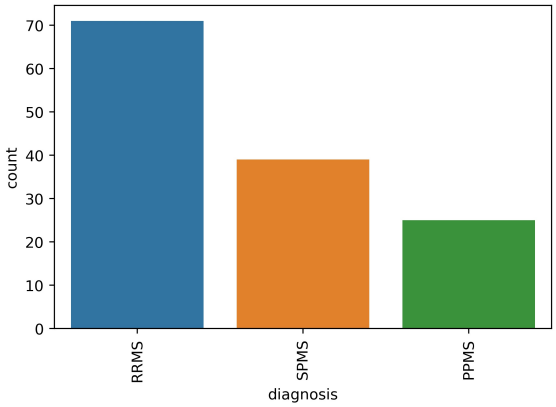


Figure 4.1: Number of patients by type of diagnosis.

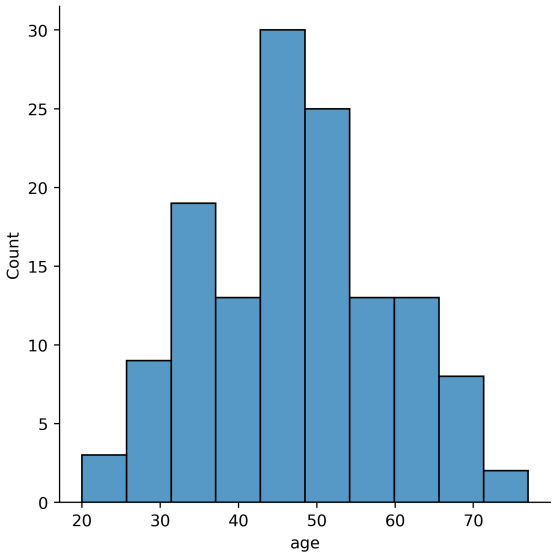


Figure 4.2: Age density

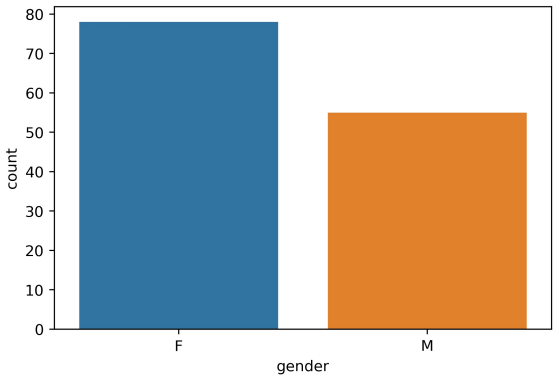


Figure 4.3: Patient count per gender

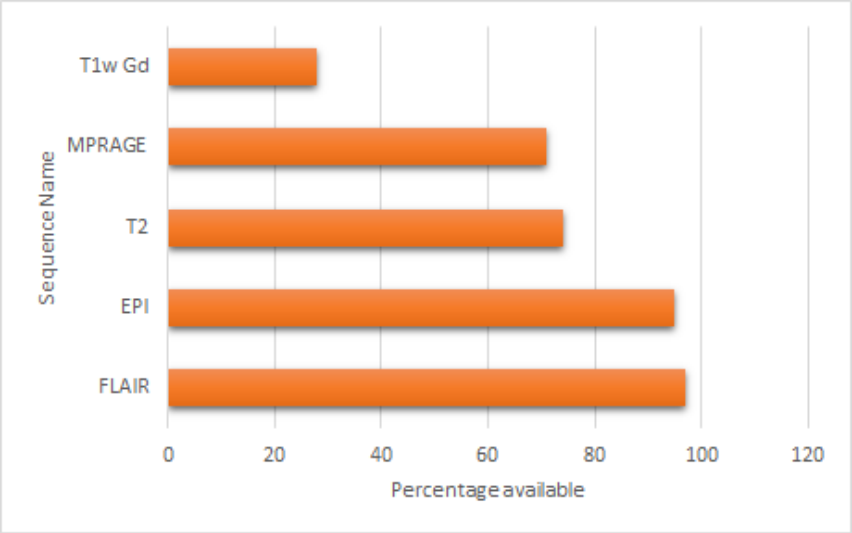


Figure 4.4: Percentage of patients for which each sequence was acquired and for which there were no technical problems.

Since i) the patients' MR images were acquired in 3 different centers and ii) technical problems were not uncommon, we did not have the same set of MRI contrasts for every patient at the end. Figure 4.4 shows the state of the image database.

4.2 Image processing: extracting MRI features

4.2.1 Convert DICOMs to BIDS format ("*Dicom2BIDS*")

The BIDS format (cf section 3.1.2), being a very convenient way to organize brain imaging data, is expected by every tool or script I wrote during my thesis. However, as medical researchers tend to be less familiar with computer tools, I developed a tool that would convert raw DICOM series freshly extracted from the MRI software directly into the NIfTI/json and BIDS format.

This tool is actually a pipeline of Python functions callable in a Jupyter Notebook combined with some other utility functions. Depending on the user's needs, he/she is able to only run parts of the pipeline or the pipeline as a whole. Hereafter you will find all the steps of this pipeline, as well as some utility functions proven to be quite useful when creating or handling a BIDS formatted database. The whole code has been made accessible on github[1] and is regularly updated to fix bugs or add new features.

Pipeline

1. The first step of the pipeline consists of finding all DICOM files in the folders and/or sub-folders of a given input folder and converting them all to NIfTI

[1]https://github.com/maxencewynen/MSdicom2bids

format using *dcm2niix* with the anonymisation option. According to each software/machine that was used to acquire the images, these raw DICOM images can be organized and/or named differently. For example, the General Electric machine of the Cliniques Universitaire Saint-Luc (CUSL) outputs a lot of subfolders each containing the DICOM files of a specific sequence of the protocol, while Erasme Hospital's *Ingenia* (manufactured by Philips) outputs a single folder with all the protocol's DICOM files mixed together. This step of the pipeline works on every data it ever has been run on (which means all data used by the lab), but it probably needs to be tuned in order for it to work on data coming from other MRI machines/softwares. It used functions from the *os*, *pydicom* and *subprocess* python libraries.

2. The second step consists of creating all directories needed for a specific pair of subject ID - session ID. The user may or may not specify one or both of these information. In the case where the user does not specify the ID, the script automatically deduces its value from the database.

3. The third and final step of the pipeline consists of moving and renaming every useful newly-converted NIfTI file into the right location. To achieve this, the function looks at the 'Series Name' in the json file accompanying every converted NIfTI file and tries to match it with a database of Series Names linked to their BIDS names.[2] This step is the most limiting one in the whole pipeline, since every MRI center (and often every research group as well) has their own naming habits for a sequence.

Utility functions

Having both worked a lot with BIDS databases and observed people working on them, I noticed most of the work done manually was very tiresome, time-consuming and favorable to making mistakes. The following features added to the Dicom2BIDS tool aim to save time to researchers and avoid some common mistakes:

- Deleting a subject from the BIDS database
- Deleting a particular session from a particular subject
- Renaming a subject, i.e. giving it a different ID
- Reorienting an image along a certain axis

Jupyter Notebook

I chose to use a Jupyter Notebook as a frame to embed all utility functions I wrote for two reasons. As the people using this script would mainly be people in the Multiple Sclerosis Image Processing Lab of UCLouvain, and as this team is currently

[2]For example, if the NIfTI file's Series Name is "3D_FLAIR", it will rename the NIfTI file as well as the json file associated to it as "sub-xxx_ses-yy_FLAIR" (where xxx represents the subject ID and yy the session ID).

only made of clinicians, the first reason was to make it accessible and not too difficult to handle for them. The other reason is to hide away all important code from the user. A further improvement could be to make a Graphical User Interface (GUI) to make it even more straightforward to run.

4.2.2 Lesion Volumetry and Localisation ("*les_voloc*")

Once we have an image database in BIDS format, we can start extracting features from the images themselves. In particular, brain volumes (including lesion volume) and lesion localization are basic yet very important features we can extract from the images. As stated in the theoretical part of this thesis, MPRAGE is the go-to sequence when doing brain volumetric analysis, while FLAIR images are the most useful to segment lesions in MS.

During the early stages of my thesis, I wrote an automated pipeline (I called "*les_voloc*", by the name of the bash command associated to it) based on FreeSurfer's recon-all, SAMSEG, bash and Python scripts. This pipeline originally takes an MPRAGE and a FLAIR image of a patient as input and outputs brain volumes of interests[3], the total lesion volume, as well as a lesion-based database including the volume and location for each lesion for that patient. Moreover, since a lot of patients included in the project lacked the MPRAGE, an option has been added a posteriori to handle patients for which only a FLAIR was available (even if the volumes other than the lesion volume were not taken into account, since segmenting brain structures based only on a T2 contrast is prone to errors). This pipeline was rarely used as a whole, because the FreeSurfer step took too much computational and manual time. However, since all parts of this pipeline were used during my thesis (and also frequently by medical researchers for their own projects), I find interesting to include the whole project (which also has a Github repository[4]).

The different step of *les_voloc* are (Figure 4.5 shows the whole flowchart):

0. **Step 0:** A prerequisite of this pipeline is having run FreeSurfer's recon-all beforehand. This would have output a 3D brain segmentation file called *aseg.mgz*. More on how FreeSurfer works can be found on their wiki [5].

1. **Steps 1-2:** These steps prepare the MPRAGE and FLAIR images for SAMSEG. In particular, step 1 performs FreeSurfer's motion correction on the FLAIR, which results in a standardized 255 cubed image with 1mm isotropic voxels and a reduced size in terms of memory space. For simplification, we will refer to this operation as a normalization (although it is more of a standardization). Step 2 does the same with the MPRAGE before registrating it to the normalized FLAIR image.

2. **Step 3:** Step 3 consists of running SAMSEG on FLAIR (or on MPRAGE and FLAIR) images. As explained in section 3.1.4, SAMSEG outputs among

[3]Intracranial, gray matter, white matter, thalamic, putamen, caudate and ventricular volumes.
[4]https://github.com/maxencewynen/Lesion-volumetry-and-location
[5]https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferWiki

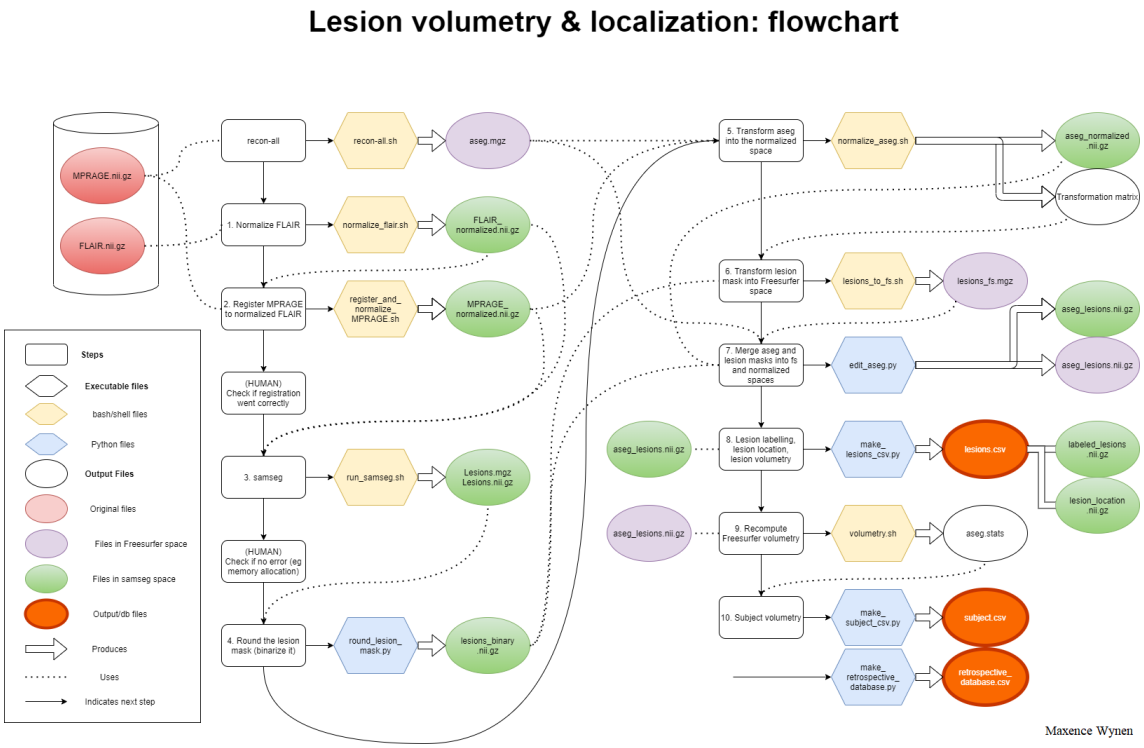


Figure 4.5: *les_voloc* complete flowchart

others a segmentation volume called *seg.mgz* and a file including the posterior probability for each voxel to be a lesion.

3. **Step 4:** Step 4 transforms the lesion probability mask into a mask of 1's and 0's according to a certain threshold. This threshold is set by default to 0.5 but can be chosen by the user according to their needs (knowing a lower value would consider more voxels to be lesions).

4. **Steps 5-6:** These steps assure (by a set of transformations and registrations) that both the aseg.mgz file and the output files of SAMSEG are on the same coordinate system. It produces two files: the FreeSurfer segmentation in the 'normalized' (SAMSEG) coordinate system and the SAMSEG segmentation in the FreeSurfer coordinate system.

5. **Step 7:** This step merges the segmentations produced by FreeSurfer and SAMSEG in each original coordinate system. In particular, it keeps the FreeSurfer segmentation but sets every voxel indicated by SAMSEG as being a lesion to the 'lesion' label.

6. **Step 8:** Step 8 gives a lesion ID to every connected components of the lesion mask (i.e. lesion labelling) and computes for each lesion its volume in $1mm^3$ voxels and its location (more on lesion localization below). This step outputs a 3D labeled lesion mask, a 3D mask of each lesion's location and a *.csv* file proper to the patients including for each lesion its volume and location.

7. **Steps 9-10:** re-compute the brain tissue volumes using FreeSurfer and the updated segmentation volumes after having "filled" the lesions (more on lesion filling in the next paragraphs). These steps output a *.csv* file for the patient with the following columns:

   - Subject ID
   - Total Intracranial Volume (TIV)
   - Normalized[6] brain volume
   - Normalized white matter (WM) volume
   - Normalized gray matter (GM) volume
   - Normalized ventricles (including all four ventricles) volumes.
   - Normalized thalamic volume
   - Normalized putamen volume
   - Normalized caudate volume
   - Normalized cerebellar GM, WM and total volume
   - Total lesion volume
   - Total number of lesions
   - Percentages of lesions per location

   This organization makes it straightforward to merge all patients' files together into a single 'patient-based' database.

Lesion localization

As already stated in Chapter 2, the locations of a patient's lesions are very important in both the diagnosis and the prognosis of MS. In this work, we identify 4 types of lesions based on their location: deep white matter lesions, periventricular lesions, juxta-cortical/cortical lesions and infratentorial lesions. To do so, we dilate[7] the structure of interest by a fixed number of iterations (an illustrative example of dilation is shown on Figure 4.6). Once this is done, we count the voxels matching both the lesion segmentation and the brain structure segmentation: if the volume of the matching voxels exceeds a certain percentage of the whole lesion volume, we then consider the lesion to be next to that specific brain structure.

SAMSEG

As already mentioned in section 3.1.4, I chose to discard the FreeSurfer's recon'all segmentation for several reasons. This decision was however taken well after the pipeline was written. To obtain the same information I would have had if I had run the whole pipeline, I rewrote parts of this pipeline such that it would work without the FreeSurfer part and outputs and only with SAMSEG. The methodology is however very similar as what has previously been described.

[6] A normalized volume corresponds to the original volume divided by the TIV.
[7] https://www.wikiwand.com/en/Dilation_(morphology)

(a) (b)

Figure 4.6: Example of a lesion segmentation before (a) and after (b) two iterations of dilation.

Lesion filling

Because an automatic brain segmentation tool such as Freesurfer has been designed to work on non-lesioned brains (healthy patients, Alzheimer, Dementia, ...), the lesions found in MS were a source of errors. [77] The solution to this problem is called *lesion filling*, and consists of replacing the lesion voxels by a certain value close to the WM intensity value. While several techniques exist, [78][79] I chose the one integrated to FreeSurfer. This was done by forcing the lesion segmentation found through SAMSEG onto the segmentation computed with Freesurfer.

Impact

Three abstracts were written based on this pipeline, using either only parts of it or as a whole. The first[11] was a technical abstract (that will be detailed in section 4.2.4) presented at the ISMRM[8] annual committee in May 2021. The second, concerning the effect of a disease modifying treatment on paramagnetic rim lesions in MS was sent to the European Academy of Neurology (accepted) and ECTRIMS[9] (a more elaborated version) this year.

The last one, based on the database we constructed during this thesis, has just (May 2021) been submitted to ECTRIMS and investigates the association between inflammatory and cerebral small vessel disease related clinical/MRI features and centrum semiovale expanded perivascular space. Moreover, this tool is used by a 1st year medical PhD student on a weekly basis, suggesting that the pipeline is suited to anyone without a special training in digital image processing or computer science.

4.2.3 Other MRI features

The best way to save time in MRI feature analysis in MS would probably be to automatically and accurately assess them by a computer. Even if researchers' efforts

[8] *International Society for Magnetic Resonance in Medicine*, https://www.ismrm.org/

[9] *European Committee for Treatment and Research in Multiple Sclerosis*, https://www.ectrims.eu/

tend to go that way ([10], [9] to cite only a few), this is unfortunately not the case yet: many MRI features still have to be assessed by visual inspection performed by neuroimaging scientists (often medical doctors, radiologists and neurologists). Below is an exhaustive list of this thesis's database MRI features obtained through expert's visual inspection.

- Presence of spinal cord lesions, identified by Dr. Valentina Lolli and neuroradiologists of Erasme, Saint-Luc and CHUV
- Presence of gadolinium-enhanced lesions, identified by Dr. Valentina Lolli and neuroradiologists of Erasme, Saint-Luc and CHUV
- Percentage of perivenular lesions (lesions with Central Vein Sign), identified by Dr Pietro Maggi, Dr. Valentina Lolli and Dr François Guisset.
- Number of paramagnetic rim lesions (PRL or rims), identified by Dr. Pietro Maggi and Dr Martina Absinta.

4.2.4 Longitudinal assessment of paramagnetic rim lesions using RimNet

Model	Session	TN	FP	FN	TP	Accuracy	F1	TPR	FPR	PPV	NPV
PHASE + FLAIR*	1	**198**	26	**14**	**40**	**85.6**	**66.6**	**74.0**	**11.0**	60.0	**93.0**
	2	**337**	**35**	**16**	**37**	**88.0**	**59.2**	**69.8**	9.4	**51.3**	**95.4**
PHASE + FLAIR	1	205	**19**	41	13	78.4	30.2	24.0	8.5	40.6	83.3
	2	332	40	26	27	84.5	45.0	50.9	**10.7**	40.2	92.7
PHASE + T2*	1	203	21	21	33	84.8	61.1	61.1	9.3	**61.1**	90.6
	2	334	38	22	31	85.8	50.8	58.5	10.2	44.9	93.8

Figure 4.7: RimNet Longitudinal: Complete metrics per session for each model tested. Abbreviations: TN, True Negatives; FP, False Positives; FN, False Negatives; TP, True Positives; TPR, True Positive Rate; FPR, False Positive Rate; PPV, Positive Predictive Value; NPV, Negative Predictive Value. Adapted from [11].

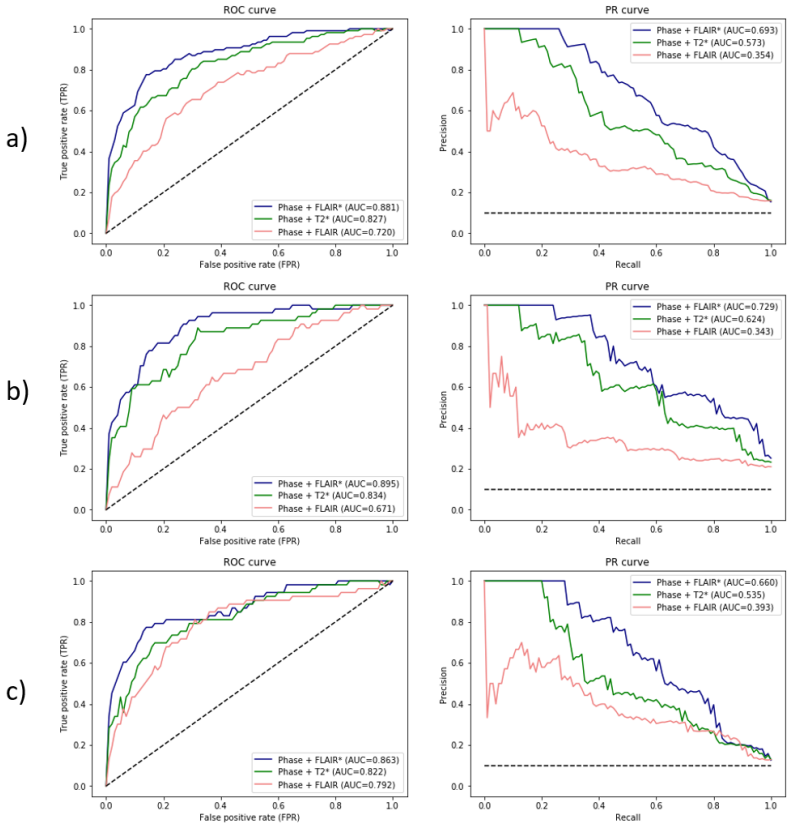


Figure 4.8: RimNet Longitudinal: Combined (a) and session-based (baseline (b) and follow-up (c)) receiver-operating characteristic (ROC) and precision-recall (PR) curves per model tested. AUC: area under the curve. Adapted from [11].

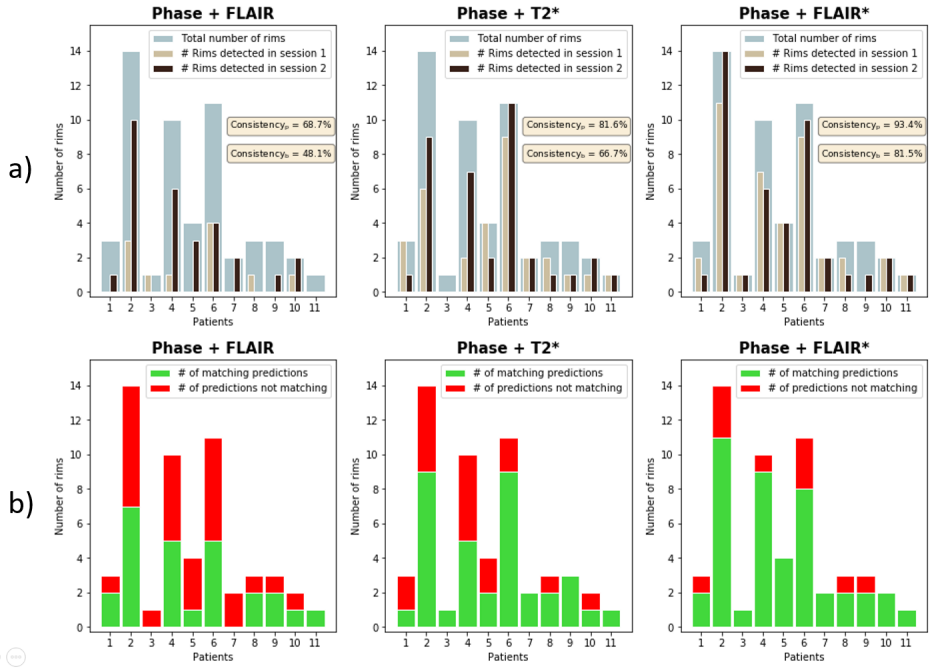


Figure 4.9: RimNet Longitudinal: Summary of the lesion-wise results for each patient with PRL. a) Number of rims detected in session 1 (baseline) and session 2 (follow-up) for each patient, overlaid on the total number of rims adjudicated in the consensus manual reading. b) Number of matching predictions between baseline and follow-up. Abbreviations: Consistency$_p$, probability-based consistency; Consistency$_b$, binary-based consistency. Adapted from [11].

During this year, I had the chance to be in the frontline of a project in close collaboration with researchers from several renowned universities worldwide. This project aimed to assess RimNet's longitudinal performance along with its generalizability and ended in a published abstract for which I was the first author [11]. As a reminder, RimNet[10] is a 3D U-Net inspired Convolutional Neural Network (CNN) developed for automated detection of PRL in MS using multiple MR contrasts acquired at 3T.

In this abstract, two experts analyzed the baseline and followup images of 13 progressive MS patients from 2 different centers (and MRI manufacturers), by manually annotating the PRL. For each of the patients, the MRI protocol included a 3D FLAIR and a 3D T2*-EPI. With those two sequences, we i) extracted the T2*-phase and T2*-magnitude of the 3D EPI and ii) generated a FLAIR* by multiplying the FLAIR and the EPI magnitude images voxel-wise. Pre-processing of the images included the lesion segmentation performed by SAMSEG, the rigid registration of the lesion mask and all other images to the T2* coordinate space, the skull stripping performed by FSL [80], and the normalization of the images according to two reference images by histogram matching[10]. Moreover, images acquired at the follow-up were rigidly registrated to the baseline. We then proceeded through RimNet to finally extract 3D lesion patches.

[10]More on histogram matching: [81]

We used two pre-trained models of RimNet for this abstract: one trained on FLAIR and phase images (phase+FLAIR model), and the other on T2*-magnitude and phase images (phase+T2* model). In particular, we tested the phase+FLAIR model on phase and FLAIR images; but the phase+T2* model was tested not only on phase and T2* images (phase+T2*) but also on phase and FLAIR* images (phase+FLAIR*). As a result, RimNet achieved an accuracy and a Precision Recall (PR) Area Under the Curve (AUC) close to the one found in its original paper (88% v 94%; 69% v 70%). Figures 4.7 and 4.8 display the detailed results. Moreover the abstract also aimed to measure its consistency. This was done using the following formula:

$$c = 1 - |p_1 - p_2|$$

, where c is the (binary or probability) consistency and p_1 and p_2 are respectively the binary or probability predictions of RimNet for the baseline and the followup. The results found were that RimNet had a binary consistency (after applying a threshold) of 82% and a probability consistency of 93% (see Figure 4.9).

4.3 Clinical features

This database's goal was to combine information collected by MRI with clinical data. Hereafter you will find a list of all clinical features explored.

- Expanded Disability Status Scale (EDSS)
- MRI acquisition date
- EDSS Date: Date on which the EDSS has been assessed
- Clinical site
- MRI site
- Birth Date
- Gender
- Date of first MS related symptoms
- Date of MS diagnosis
- Ongoing treatment: 11 identified treatments + no treatment
- Start of ongoing treatment (Date)
- Date of last relapse (if there was a relapse)
- Type of last relapse (if there was a relapse)
- Presence of corticosteroid after the last relapse (if there was a relapse)

- Number of relapses
- Start of progression (if the patient is diagnosed with progressive MS)
- Arterial hypertension
- Hypercholesterolemia
- Smoking
- Diabetes
- Obesity
- Presence of specific oligoclonal bands in the cerebrospinal fluid (CSF)

4.4 Multi-modal database discussion

After almost a year of working on this project, I noticed that data - both MRI and clinical - can easily become unorganized. This was especially the case when the research group grew bigger. Without any good practice, the lack of data organization lead to loss of information, mislabeling data, or simply a huge waste of time. This is why I immediately started organizing the images I received in a BIDS format.

Nonetheless, even if the BIDS standard is very convenient when working in a small lab with a few people, it is not a sustainable database management system (or even a database management system at all) and has a few flaws when comparing with the way one has to correctly store data. One issue is that it mainly handles images and some really basic information on patients included in the studies, whereas our studies (and this thesis shows a perfect example of it) make use of both clinical and imaging data. Another example of an issue is that a BIDS database can hardly be shared by two (or more) researchers of the same lab: everyone has to keep their own version of the database on their local storage.

This section aims at first establishing the specifications for a database management system (DBMS) for a multi-modal database including several kinds of medical images as well as clinical information, then explores a few existing DBMS that could fit to the specifications.

4.4.1 Specifications

A certain number of factors are to be considered when choosing the right DBMS for a project[11]. Hereafter we will discuss some of them by defining which attributes such a multi-modal DBMS in MS should have.

[11]Some of these factors were inspired by https://datahq.co.uk/knowledge-hub/blog/top-10-considerations-when-choosing-a-database-management-system

- **Usability**: The usability describes which user profiles will have to interact with the system. In our case, the users will be of two types: either medical researchers with little to no knowledge of computer science or engineers who should be very comfortable with it. All users should be able to both add, retrieve, inspect and visualize data easily.
- **Security**: Security of the data is paramount when considering a DBMS handling confidential medical data. Moreover, it should both avoid the physical risk of losing data and the unintentional corruption of data by human error.
- **Integration:** All actions available to a user must have the option of being processed automatically, i.e. to be easily integrated into automatic pipelines. In particular, a user should be able to export the image database to a BIDS format.
- **Scalability:** The 3D images (and particularly when in DICOM format) take up a lot of space in memory. After investigation on a sample of 41 patients, a DICOM directory of one MRI session in MS takes an average of 992 MB ($\pm$420 MB, range [93 ; 2330]), while the associated space in the BIDS directory[12] is on average 671 MB ($\pm$136 MB, range [292 ; 882]). The details of these computations can be found in Appendix A.
- **Visualisation:** Visualisation should be non-developer user friendly.
- **Functionality:** The ideal DBMS should offer some very useful features such as:
  - Data extraction and filtering
  - Data visualisation and reporting
  - Automation
  - Insight and analysis
  - Data exporting
  - Easy integration to a bigger database
  - ...
- **Support:** Any engineer should be able to keep the database updated, even years after it has been developed. The existence of a User Manual, both for basic users and developers, is crucial.

4.4.2 Exploration

Existing international multi-modal databases in MS

To the best of my knowledge, there are only two international multi-modal databases projects in MS as of today. The first one, called *MSBase*[13], was originally

[12]Including all of the images derivatives, like the transformations, registrations, segmentations, ...
[13]https://www.msbase.org/

Australian but grew out to be international. This database relies on neurologists/neuroradiologists to fill in the clinical and radiological information of the patients they want to include via a program called iMed. The second one includes the latter and is called Big MS Database[14]. While these databases look very useful for basic research, they do not integrate (yet) more advanced research features in MS (like PRL, CVS or multishell diffusion). Another interesting international multi-modal database project in medicine to mention is the Alzheimer Disease Neuroimaging Initiative (ADNI). Unfortunately, I have not been able to find any technical papers explaining which DBMS each of these projects use.

Types of database management systems

When choosing the right DBMS, we always want to look at what we may call "traditional" database applications, i.e. storing and handling only textual and/or numeric data, relying on the relational model (such as SQL). A solution to still be able to handle images would then be not to store them in the database, but rather storing the path to where to find them. Traditional databases are, however, not the only existing DBMS. During the last two decades and with the emergence of social media such as Facebook or Twitter, a whole new kind of DBMS called NoSQL has been developed. NoSQL DBMS, interpreted as "not only SQL" rather than "no to SQL", are generally conceived to store huge amounts of data without necessarily having a relational based structure.

Before looking at the different categories of NoSQL systems, we should fully understand the **CAP Theorem**. This theorem explains the competing requirements in a distribution system with replication. CAP stands for *Consistency* among replicated copies, *Availability* of the system when performing read and write operations, and *Partition tolerance*. The theorem states that it is impossible to guarantee all three of these properties at the same time in a distributed system with data replication. Contrary to the very consistent SQL applications[15], NoSQL systems tend to give more importance to Availability and Partition Tolerance aspects than to Consistency. Based on these properties, we may now develop the 4 categories of NoSQL systems [82] (cf Figure 4.10 for a concise view):

1. **Document-based NoQSL systems**: Data is stored in *documents* using well-known formats like JSON. Documents can be accessed either with their document id or by other indexes. Notorious examples of document-based NoSQL systems include MongoDB and CouchDB.

2. **NoSQL key-value stores:** Each key is associated to a value that can be any complex data structure. These simple systems make the access to data very fast. Amazon's DynamoDB is a famous example of a NoSQL key-value store.

[14] https://bigmsdata.org/

[15] See the ACID properties of relational DBMS.

3. **Column-based NoSQL systems** stores data in columns rather than by rows. It is particularly useful in the case of horizontal scaling. Hadoop Distributed File System is a nice example of column-based NoSQL systems.

4. **Graph databases:** Graph databases like Neo4j stores data in a graph, which is a collection of vertices (nodes) and edges.

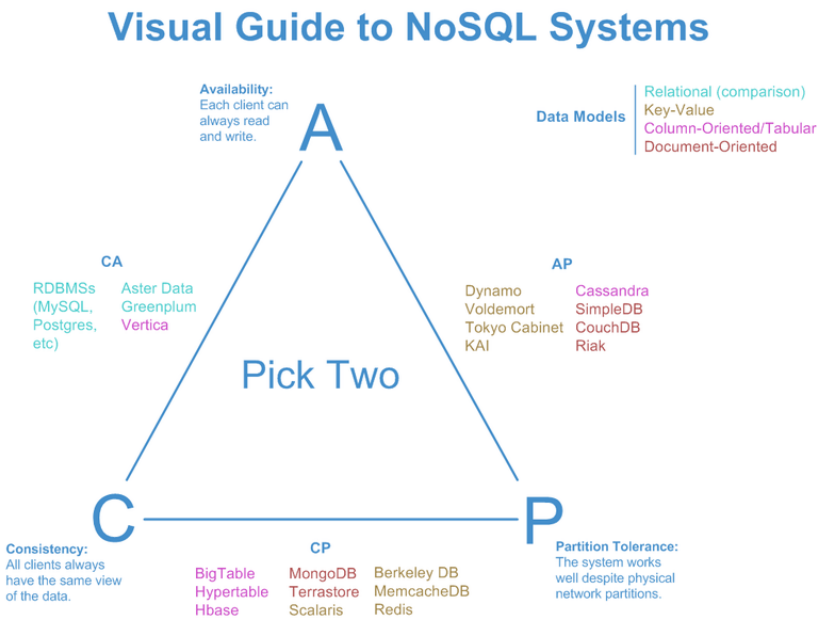


Figure 4.10: The CAP theorem illustrated. Source: https://bohutskyi.com/cap-theorem.html

In my opinion, the best DBMS for this project should be a NoSQL system, and particularly a document-based system like MongoDB. Arguments for this choice include the fact this type of DBMS handle images and metadata very easily and are very flexible and scalable.

4.5 Data cleaning

Before being able to perform any statistical analysis or to apply any machine learning model to the database, I first had to pre-process it. This pre-processing consisted of several steps:

1. First I had to change all date-based features into numerical features. To do so, I took as reference the MRI acquisition date and computed the difference in months (except for the *age* feature) with it. This leaves us with either positive or negative values, corresponding respectively to dates before and after the reference MRI acquisition date.

2. The second step consisted of labelling all categorical features to numbers. This is a usual step when pre-processing data in order to perform statistical analysis on it.

3. For the third step, I have augmented the data by adding features based on other features. In particular, I computed:

    - the Multiple Sclerosis Severity Score (MSSS) based on the EDSS and the disease duration
    - the Age-Related Multiple Sclerosis Severity (ARMSS) based on the EDSS and the age
    - the number of cardiovascular risk factors
    - a categorization of patients based on the number of PRL: 0 PRL, [1-3] PRL and more than 4 PRL. This is supported by the recent literature on this subject. [6][5]

4. The last step of preparing data for statistical analysis was removing irrelevant features or features with large proportion of missing data. For the latter, I allowed a maximum of 30% missing data per feature. Indeed, only 70% of the patients had an MPRAGE image. Because MPRAGE is the go-to sequence for brain tissue segmentation (cf section 2.2.5) and extensively for lesion location, the same cutoff of missing data was applied for all features. Hereafter is a list of all dropped features with the reason why they were dropped:

    - All dates (irrelevant and unnecessary since it has already been transformed).
    - Clinical and MRI site (irrelevant)
    - All clinical information related to the last relapse
    - Presence of spinal cord lesions
    - Number of lesions with the CVS per location
    - Elapsed time since the start of the treatment

Chapter 5

Exploratory and statistical analysis

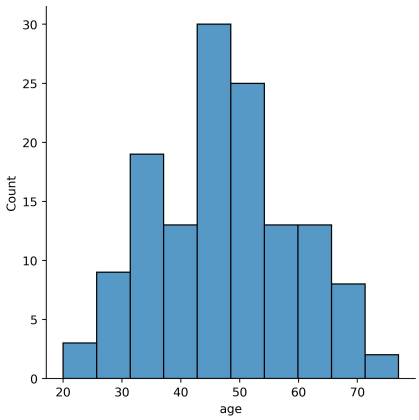
Once all data had been gathered, processed and cleaned, we ended up with a database well suited for statistical analysis. We still had missing data however, but we were already able to perform a lot of tasks and gain some knowledge on the data without any imputation. In particular, we could explore the data by describing it, analyzing the distributions, test for their normality etc. We could also perform some simple bivariate analysis that may either invalidate, confirm or reveal some insights on our data.

5.1 Data exploration

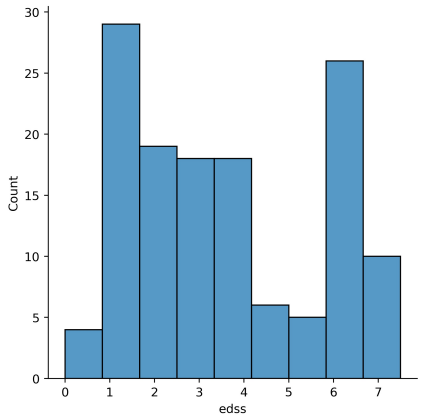
I started by computing simple statistics. Those included the mean, the median, the minimum and maximum values, the standard deviation and finally the skewness[1] of every quantitative variable. Table 5.1 displays these measures for each of one of the quantitative variables in the database. Additionally, I drew the distribution plot of each of these quantitative variables. A small selection of distribution plots can be found on figure 5.1. For what concerns the qualitative variables, I chose to illustrate them with a histogram counting the occurrence of each value that the variable could take (Figure 5.2).

The average age of the patients in our database is 47 ($\pm$ 11.9) years. This number is consistent with the fact that the average age at onset in MS lies between 20 and 40. An average age of 47 means that most of the patients have suffered from MS for a long time (several years). This is also confirmed by the average disease duration (discarded feature because of the lack of data) being almost 15 years, and the average disability/severity scores being relatively high (around 4). With this information, I was able to compare the different brain volumes with the literature. Indeed, many studies have shown a strong effect of aging on brain atrophy. [83] The average total intracranial volume (TIV) computed on our sample is in line with the literature, as well as all the other brain volumes (confirmed by [84] and expert MS clinicians

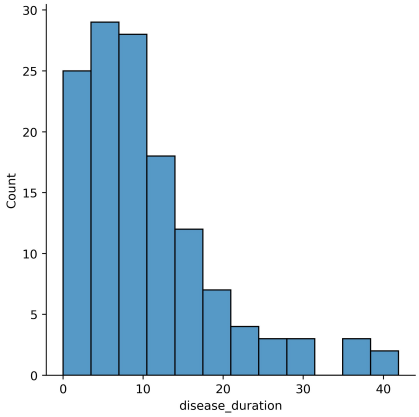
[1] The skewness is a measure of the asymmetry of a continuous variable's distribution around its mean.



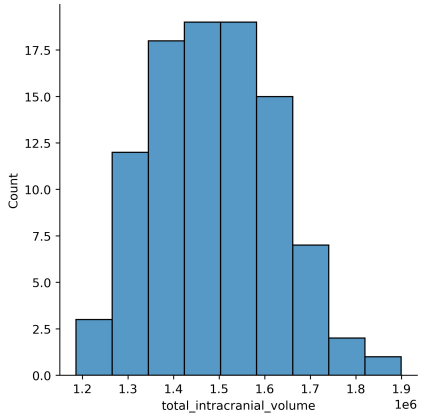
(a) Age distribution



(b) EDSS distribution



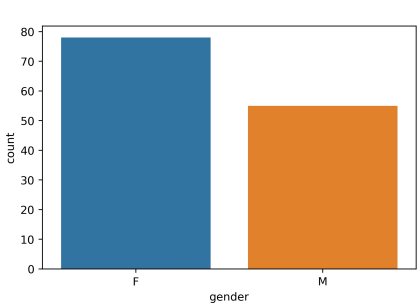
(c) Disease duration distribution



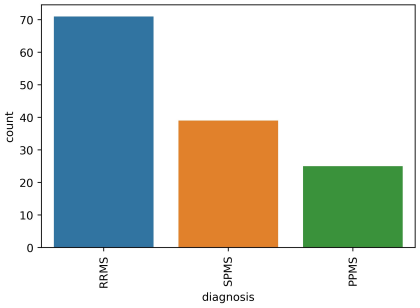
(d) Total intracranial volume distribution

Figure 5.1: Examples of histogram distribution of quantitative variables

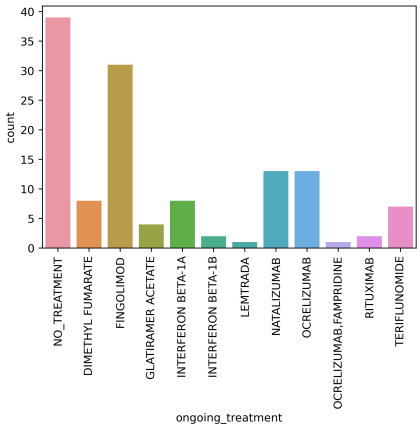
(Dr Pietro Maggi Dr Valentina Lolli)). These results bring further support for the methods I chose to measure the brain volumes.



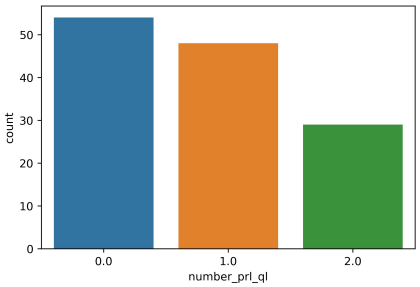
(a) Gender



(b) Diagnosis



(c) Ongoing treatment



(d) Number of PRL (0: 0; 1: [1-3]; 2: ≥ 4)

Figure 5.2: Examples of count plots for qualitative variables

	Mean	Median	Minimum	Maximum	Standard deviation	Skewness
EDSS	3,6	3	0	7,5	2,1	0,343
MSSS	4,8	4,3	0,2	9,8	2,8	0,183
ARMSS	4,8	4,5	0,6	9,5	2,6	0,154
Age	47,3	47	20	77	11,9	0,114
Total intracranial volume (mm^3)	1490054,5	1483911,5	1186110,7	1898970,6	137329,4	0,3
Normalized cortical volume	0,3169	0,3165	0,2397	0,3643	0,0279	-0,324
Normalized white matter volume	0,2525	0,2566	0,1833	0,294	0,0227	-0,519
Normalized brain volume	0,5694	0,5747	0,4231	0,6334	0,0451	-0,648
Normalized thalamic volume	0,0076	0,0077	0,0052	0,0094	0,001	-0,446
Normalized caudate volume	0,0045	0,0046	0,0019	0,0057	0,0008	-0,738
Normalized putamen volume	0,007	0,0069	0,0042	0,0103	0,0011	0,534
Normalized ventricular volume	0,0202	0,018	0,0096	0,0724	0,0094	2,616
Normalized cerebellum gray matter volume	0,0665	0,0665	0,0512	0,0893	0,0071	0,456
Normalized cerebellum white matter volume	0,0162	0,0164	0,0108	0,0203	0,0019	-0,26
Normalized cerebellum total volume	0,0827	0,083	0,0645	0,1096	0,0087	0,342
Normalized total lesion volume	9112	5088	0	60476	10468	1,832
Number of deep white matter lesions	5,2	4	0	19	4,4	1,131
Number of periventricular lesions	6,1	6	0	26	4	1,455
Number of juxta-cortical or cortical lesions	3,6	3	0	14	3,2	1,437
Number of infratentorial lesions	0,7	0	0	6	1,1	1,92
Total periventricular lesion volume	7684,5	3253,5	0	36812	9261,3	1,346
Total deep white matter lesion volume	911,5	350	0	36074	3647,3	9,271
Total juxta-cortical or cortical lesion volume	226,8	121	0	1908	288,8	2,777
Total infratentorial lesion volume	64,1	0	0	1175	164,9	4,436
Elapsed time (days) since the EDSS evaluation	7,5	-4	-485	729	123,5	1,307
Percentage of perivenular lesions	70,1	73	23	100	17,3	-0,473
Number of cardiovascular risk factors	0,96	1	0	5	1,24	1,494

Table 5.1: Basic statistics of quantitative variables.

An investigation was also performed on the other (qualitative) features present in the database. The female/male proportion is approximately equal to 1.4, which means that there are approximately 1.4 times more females than males. This proportion is in line with the literature. [85] As regards the different cardiovascular risk factors, I compared the proportions computed in our database with the overall population. The world health organization[2] (WHO) estimates the number of people worldwide affected by arterial hypertension to 1.13 billion, which corresponds to a proportion of roughly one person out of 7, with a smaller number for women compared to men. In comparison, our sample has a proportion of 12.5% of patients with arterial hypertension (between 1/8 and 1/7), which can be explained by the fact that our sample features more females. The WHO also states that about 7% of the world population suffers from diabetes, and this is also the number we get when we compute this number for our database. Still according to the WHO, 15% of Belgian and Swiss citizens are obese against 9% in our database and 24% are smokers while 35% of the patients we included in the projects smoke.

Additionally to the previous analysis, I also investigated the shape of every quantitative variables' distribution, and I particularly studied their normality. A summary of the results of each quantitative variable's normality test can be found on table 5.2. Note that this analysis was done each time without considering the various sub-groups (i.e. male/female, RRMS/PMS, etc.). This is probably why most of the distributions do not seem to follow a normal distribution, except for a few variables like the age, the TIV, the normalized WM volume, or the cerebellum volume.

5.2 Bivariate analysis

After having explored each variable separately, I investigated the links between each pair of variables based on the methods explained in section 3.3.2. While it was obviously good to have a relatively high number of variables at my disposal (especially given their novelty), the investigation of the way each variable is related to all others also gave a large number of information. Therefore, I had to extract only the useful information from this huge amount of insights.

I started by studying the relations between the disability or severity scores (see section 2.2.3) and all other variables, because it was in line with the objectives of the thesis. I first noticed the age had a positive correlation of 0.4 with the EDSS, but almost no correlation with either ARMSS or MSSS. This disappearance could be explained by the fact that the latter take a duration in years (either the age or the disease duration) into account. Moreover, there was a negative correlation between the different scores and most brain volumes (total brain, caudate, cerebellum, cortical, thalamic and WM). This can be explained by the fact that patients with greater disability tend to have atrophy in some parts of their brain. [86][87] The ventricular volume however displays a negative correlation with the EDSS, because a higher ventricular volume implies a more important brain atrophy. As to the total lesion

[2] https://www.who.int/

	Shapiro-Wilk	**D'Agostino k^2**	**Normality**
age	TRUE	TRUE	TRUE
edss	FALSE	FALSE	FALSE
total_intracranial_volume	TRUE	TRUE	TRUE
cortical_volume	FALSE	TRUE	FALSE
white_matter_volume	TRUE	TRUE	TRUE
brain_volume	FALSE	FALSE	FALSE
thalamic_volume	FALSE	TRUE	FALSE
caudate_volume	FALSE	FALSE	FALSE
putamen_volume	FALSE	FALSE	FALSE
ventricular_volume	FALSE	FALSE	FALSE
total_lesion_volume	FALSE	FALSE	FALSE
number_wm_lesions	FALSE	FALSE	FALSE
number_periventricular_ lesions	FALSE	FALSE	FALSE
number_juxtacortical_ cortical_lesions	FALSE	FALSE	FALSE
number_infratentorial_ lesions	FALSE	FALSE	FALSE
total_periventricular_ lesions_volume	FALSE	FALSE	FALSE
total_wm_lesions_volume	FALSE	FALSE	FALSE
total_juxtacortical_ cortical_lesions_volume	FALSE	FALSE	FALSE
total_infratentorial_ lesions_volume	FALSE	FALSE	FALSE
basal_ganglia_pvs	FALSE	FALSE	FALSE
total_pvs	FALSE	FALSE	FALSE
lacunes	FALSE	FALSE	FALSE
wmh	FALSE	FALSE	FALSE
days_since_edss_eval	FALSE	FALSE	FALSE
percentage_perivenular_ lesions	FALSE	TRUE	FALSE
number_of_cvrf	FALSE	FALSE	FALSE
msss	FALSE	FALSE	FALSE
armss	FALSE	FALSE	FALSE
cerebellum_gm_volume	TRUE	FALSE	FALSE
cerebellum_wm_volume	TRUE	TRUE	TRUE
cerebellum_total_volume	TRUE	TRUE	TRUE

Table 5.2: Results of the normality tests performed on every quantitative variable.

volume and the number of infratentorial lesions they both display a strong positive correlation with the disability which is in line with the literature.

Additionally to the correlation with the disability or severity scores, I also investigated the links between these variables and the qualitative variables. The most notable qualitative variables were the diagnosis ($p < 0.0005$) and the number of PRL ($p < 0.001$). Indeed, patients with RRMS are generally less disabled and have a less severe disease than patients diagnosed with progressive MS (PPMS or SPMS). Moreover, as stated in section 2.2.5.1, previous studies have shown that a higher PRL burden is associated to a higher disability.

Even though the relationships between all variables are interesting, it would be too heavy to comment all of them. This is why a heatmap summarizing the correlation coefficient between each pair of variable can be found on figure 5.3. We can also plot the associations on a scatter plot (Figure 5.4. This analysis was an important step before applying multivariate analysis or even more complex models developed in chapter 6.

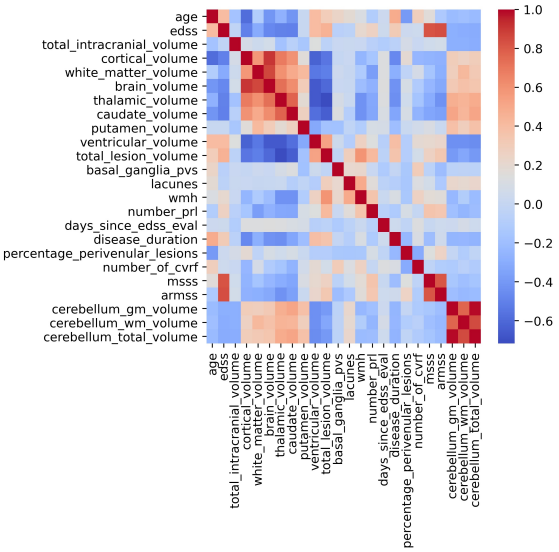
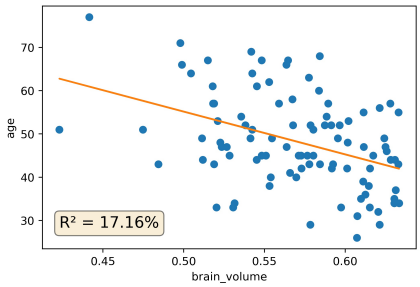


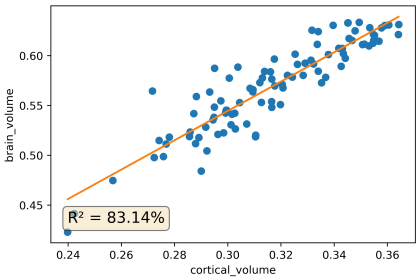
Figure 5.3: Heatmap of the correlation computed between all pairs of quantitative variables.

5.3 Additional statistical analysis

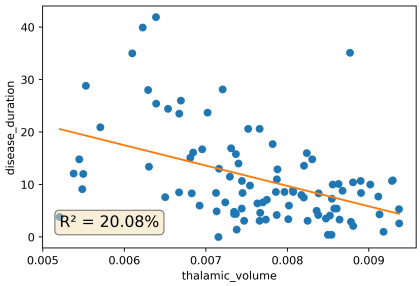
To further complete the database already created, two experts (Dr Maggi, Dr Pasi) added specific clinical and radiological features related to brain microischemia (called "cerebral small vessel disease" or CSVD). The extensive analysis of these features has provided new information concerning the coexistence of inflammatory and microangiopathic-ischemic mechanisms in MS. The addition of these variables in the statistical analysis developed in this chapter resulted in an abstract submitted to the European committee for treatment and research in MS (ECTRIMS) in May 2021.



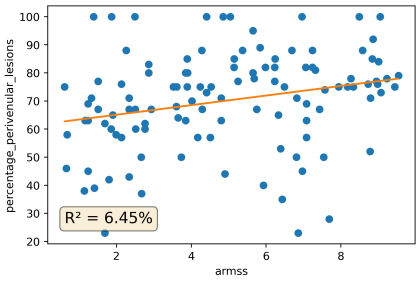
(a) Age v. Brain volume



(b) Brain volume v. cortical volume



(c) Disease duration v. Thalamic volume



(d) Percentage of perivenular lesions v. ARMSS

Figure 5.4: Linear relationships between two quantitative variables: small subset.

In particular, the abstract investigated the association between the features in the database and the score centrum semi-ovale anlarged perivascular spaces (CSO EPVS), which is a well known biomarker for CSVD. The submitted abstract suggests there is an association of a higher CSO EPVS score with age ($p<0.0001$), hypertension ($p=0.003$), a lower GM volume ($p=0.009$) and a lower percentage of perivenular lesions ($p=0.005$). It however also suggests to discard any association with a higher lesion volume ($p=0.08$) or PRL ($p=0.55$).

Additionnaly to the previously mentioned abstract, I also had the chance to do the technical parts and perform most of the statistical analysis of another abstract sent to ECTRIMS this year (2021). This abstract, adding to another previously accepted by the European Academy of Neurology (EAN) [88], studied the effect of a disease modifying treatment (DMT) on PRL. In particular, it both investigated the change in volume (in terms of Compound Annual Growth Rate or CAGR) and the change in average intra-lesion-QSM[3] value of PRL in a longitudinal cohort of 38 MS patients. The abstract concluded that the specific DMT investigated did not seem to have a large effect on PRL.

[3]As a reminder, the quantitative susceptibility mapping is a *quantitative* measure used to assess tissue integrity, see section 2.2.4

Chapter 6

Predictive analysis and clustering

The main goal of having built such an advanced multimodal database is to get new insights on the physiopathology of MS. In particular, we aim to predict the future course of the disease in singular patients (prognosis), as well as to identify different sub-types of the disease based on the available advanced features (also called the "MS spectrum", cf chapter 1). These are questions that machine learning (ML) helps answer, specifically by *regression* (for the prognosis) and *clustering* (for the segregation of the MS spectrum).

However, even if ML algorithms may approximate very complicated functions, they also require a large of amount of information to be accurate. As far as our project is concerned, however, the financial and time costs represent a major obstacle to the filling of such a database. This is a frequent issue in the field of artificial intelligence applied to medicine, and may lead to results being less accurate. Moreover, the largest difficulty we will have to deal in our case is the large amount of missing data while most ML algorithms require a complete database to work with.

This chapter will start by stating this missing values problem and how to potentially solve it, then it discusses the different approaches possible to answer to the prognosis and segregation problems. Note that by lack of time during the thesis, these steps are still in progress and no results can be shown yet. However, I plan to keep working on that analysis and publish my future results in a peer review open access journal during the summer.

6.1 Missing Data and Data Imputation

Whether because of the lack of images, poor encoding (or no encoding at all), a technical problem or some other issue, our database is unfortunately not entirely complete. Figure 6.1 shows a synthetic view of all missing data in the database, representing 15.4% of the database. As explained above and even if this is not too problematic when performing simple bivariate analysis, most ML algorithms require the database to be 100% complete. To overcome this problem, we have to use *data imputation* techniques. Data imputation is the process of replacing missing data in a database with other substitute values. The main drawback with this imputation is

that it introduces a bias to the data, which increases when the amount of data to be imputed increases as well. Consequently, the results' interpretation of any statistical or ML analysis can thus be relatively questionable if the imputation is not done well. This section investigates the causes of missing data and some practical methods to impute them.

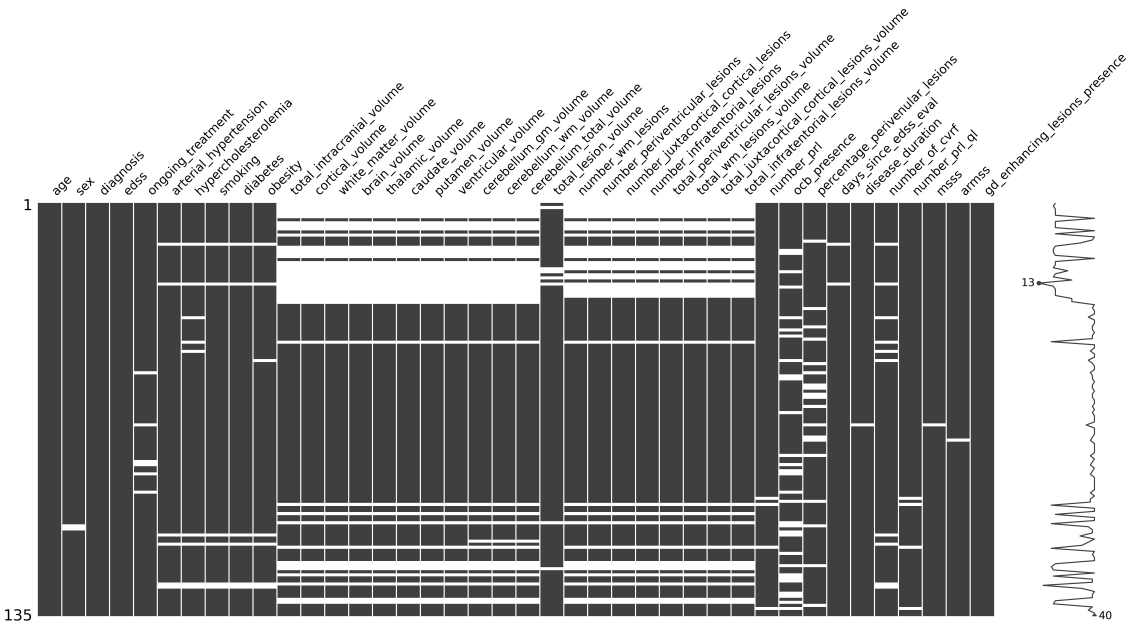


Figure 6.1: Synthetic view of missing data in the database.

The first question we want to ask ourselves is the reason why the data is missing. For example, the data sample is likely still representative of the population if values are missing totally at random. [89] However, if the values are consistently missing, the analysis may be biased. According to [90], the mechanisms underlying missing data can either be:

- **Completely at random (MCAR)**: if the circumstances that lead to any particular data-item being absent are completely random and independent on both observable and unobservable parameters of interest.
- **At random (MAR)**: if the missingness is not random, but can be fully explained by variables for which complete information exists,
- **Not at random (MNAR)**: Data that is neither MAR nor MCAR (i.e. the value of the missing variable is related to the reason it is missing), also known as nonignorable nonresponse in psychology.

The data imputation method can be different in each one of these case scenarios. In my opinion we are in the case where missing data are MAR, because although there is no underlying mechanism causing a particular data-item to be missing to my knowledge, most variables can probably be explained by the others. Hereafter you will find a non-exhaustive list of all methods available for imputing MAR data.

- **K-Nearest Neighbours**: Non-parametric method that finds the k "nearest" (i.e. most similar patients) and average their value, under the assumption that missing values can effectively approximated by the values of the points that are closest to it based on other variables. The distances used for each type of variable may be:
  - The Euclidean distance or the Manhattan distance for *numerical* variables
  - The Hamming distance [91] for *categorical* variables
  - The Euclidean distance, the Jaccard distance or the Hamming distance for binary variables
- **Deep learning based imputation methods**: includes already existing methods such as [92].
- **Regression**: This consists in finding linear or non-linear relationships between the missing feature and the other non-missing features. One of the most advanced techniques to do this is called Multiple Imputations by Chained Equations or *MICE* [93]. An extension to this well-known algorithm (SICE) has also been proposed recently. [94]

To conclude this section, let's say that the best way to deal with missing data is 'simply' to acquire (more of) them.

6.2 Predicting the disability (prognosis)

Trying to predict the future course of a patient's status is a well known problematic in MS. Since no other method really stands out, researchers started to use machine learning for this task. The main goal of this research was either to predict the future disease course i.e. trying to predict the time of conversion to SPMS in patients with RRMS, or predict the patient's EDSS. While a few approaches have already been tried based solely on MR images [95] (without using any feature extraction from the images) or solely on clinical data [96][97], there are to my knowledge no studies combining such advanced MRI biomarkers as the CVS or PRLs with clinical *and* biological data.

This prognosis can be defined as a regression problem where the target variable is either the EDSS, the MSSS or the ARMSS. Even though most researches use the EDSS as target variable in a regression framework, it would in my opinion probably be best to use scores such as MSSS and ARMSS as they are continuous, unlike the EDSS which is resembles more a qualitative scale.

There exist a lot of machine learning models that perform regression. Some, like simple multivariate regression, ridge regression or Lasso regression are linear. It is always important to try out these 'simpler' models before investigating more complex (non-linear) models, because they are a lot more interpretable. Non-linear methods include random forests or k-nearest neighbours, the latter being probably

less suited to our case since we don't have abundant data and since it does not handle the noise in the data very well. A last method I want to discuss is the support vector regression (SVR), which can either be linear or non-linear depending on the kernel used, and which is particularly powerful when having a small dataset.

6.3 Identifying different profiles of the disease

In recent years, medical doctors have relied on a specific set of criteria to diagnose MS, called the McDonald's criteria (see section 2.2.2). These criteria are essentially based on the spread of the disease in space and time. Once this diagnosis is made, the doctors will also categorize the patient's disease in one of the four accepted subdivisions of MS: CIS, RRMS, PPMS or SPMS (section 2.2.1). These categories are, however, based solely on how the disease progresses, which is potentially (and probably) not the best way to separate MS patients. In particular, the pathophysiological barriers between the different types are not well defined and the fact that certain patients respond to a specific therapy and other do not is currently largely unexplained. What criterion(s) should be used in this case? This is the whole point of this section: to try to find a way to subdivide the disease into an ideal number of parts according to optimal criteria.

This precise question has already been investigated by Dr Eshaghi and his colleagues (University College London, London, UK), who published an article in Nature.[98] In this article, the authors use pathobiological mechanisms to find a new way to subdivide the disease. More specifically, they extracted a small number of features from patients' MRIs and observed their evolution over time. This analysis, however interesting it may be, does not consider the advanced biomarkers that we have cited in section 2.2.4 nor any research sequence, but rather the sequences used in a clinical routine (T1 and T2). On the other hand, they carried out their research on an impressive number of subjects (6322), which makes their results very solid.

This problem of subdividing the disease into several sub-types can actually be solved by a clustering algorithm, i.e. an unsupervised learning method that partitions a population or set of data points into several groups (clusters) so that data points in the same group are more similar to each other and different to data points in other groups (clusters). In our case, each data point in the multi-dimensional space corresponds to a specific patient with specific features. In most clustering algorithms, we have to specify the number of clusters (often represented by the letter k). Figure 6.2 shows an example of clustering in a 2-dimensional space with k = 3.

Different types of clustering algorithms exists, each one having their own specifities[2]:

[1]Source: https://www.geeksforgeeks.org/clustering-in-machine-learning/

[2]Source: https://towardsdatascience.com/clustering-101-how-to-choose-the-right-algorithm-for-your-application-fb1521ea13fc

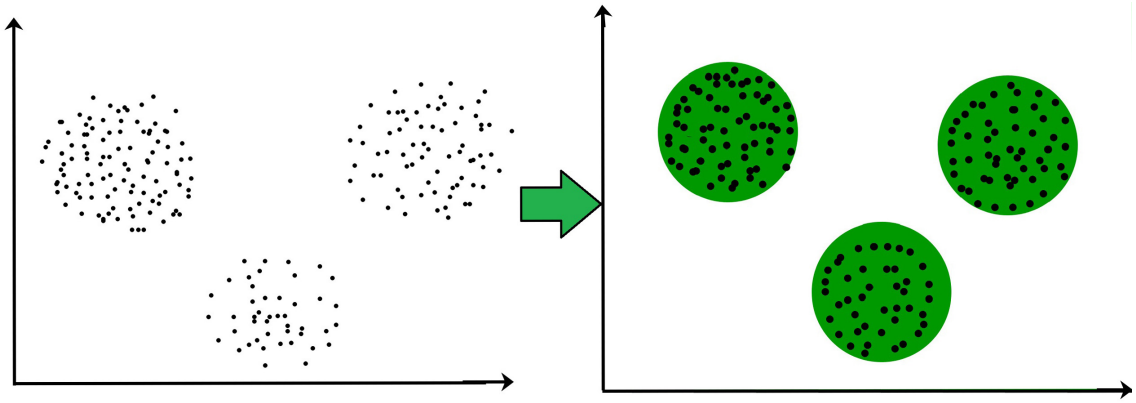


Figure 6.2: Example of the separation of data points in a 2D space into 3 different clusters. [1]

- **Centroid-based clustering algorithms**, that consider the dataset's centers as the cluster's centers. The most well-known example of a centroid-based algorithm is K-means. The perfect number of clusters may be found with the elbow method.
- **Hierarchy-based Algorithms** create a tree of clusters. This is accomplished by creating a hierarchical relationship between the various data points. At first, each data point is assigned to its own cluster, and then the two closest clusters are combined to produce a new cluster until only one cluster remains. The main advantage to using this method is that one does not need to specify the number of clusters (k) beforehand, but can rather choose a cutoff that best represents the different clusters.
- **Fuzzy Theory-based Algorithms** (also referred to as soft clustering) where data point can belong to more than one cluster. They are based on the computation of the level of belonging of each data point to all clusters.
- **Distribution-based Algorithms** assume that data points inside a cluster share the same probability distribution (often being a Gaussian distribution). Large datasets may be mined using distribution-based clustering, which yields sophisticated cluster models. As a result, the correlation and reliance between the various data points can be revealed.
- **Density-based Algorithms** create clusters based on the density of the data in a specific area. This produces arbitrary-shaped distributions in areas with a high density of points. These algorithms have trouble dealing with data of varying densities or high dimensionality.

6.4 Choosing the best model

For each of the two goals developed in this chapter (regression and clustering), a choice has to be made among all the available algorithms. Because the clustering

problem is an unsupervised learning problem and we don't really know what we are looking for, we will have to look at the results of each model and define with clinical experts which one makes the most sense. Concerning the regression however, being a supervised learning problem, we may assess more precisely which model gives the best results by splitting our data into a training, validation, and test set. For each model, we tune the hyperparameters by k-fold cross-validation to choose the best model for each algorithm. Once the best model has been chosen for all different algorithms available, we can compare their performance by applying them to the test set.

Chapter 7

Discussion

7.1 Impact

Since July and thanks to this thesis, I had the chance to join a newly established medical image analysis laboratory specialized in MS and affiliated to the UCLouvain. This laboratory is only composed of medical doctors. In addition to the work done for my thesis, I was able to help these researchers in their investigations by assisting them at a technical level, which allowed me to learn a lot about the radiological, biological and clinical basis of MS. I enthusiastically participated in the development of tools for image processing and statistical analysis resulting in a total of two accepted conference abstracts [11][88] with two other abstracts currently under evaluation. One full scientific paper on the same topics (with me as the second author) is currently under preparation.

I was the first author of an abstract accepted to the 2021 ISMRM[1] annual meeting, investigating the longitudinal performance of RimNet along with its consistency (see section 4.2.4). This abstract introduced the problem of automated longitudinal detection of advanced biomarkers in MS. The reason why this issue is important in the MS imaging research field is that many studies aim to assess the disease evolution over time using MRI. On the other hand, using longitudinal information can potentially improve biomarker detection. This principle has been applied by the developers of SAMSEG for example, who have added an extension to their program allowing to use multiple MRI sessions to improve their segmentations. During this abstract, I used the les_voloc pipeline to automate the lesion segmentation before providing this segmentation to RimNet.

I also participated into 2 other projects which resulted in 2 accepted conference abstracts (with me as second author)[11][88]. The first one studied the association between lesion type (PRL vs non-PRL) and longitudinal lesion volume or lesion QSM evolution, in patients treated or not with a specific highly effective treatment for MS. Here "les_voloc" pipeline was also used to characterize brain and lesion volumes. I helped in computing lesion's mean magnetic susceptibility

[1]International Society for Magnetic Resonance in Medicine

analysis using Xu Li's QSM program[2] and I performed most of the statistical analysis.

The second project, which was directly based on the multimodal database established during this thesis, studied the association between microangiopathic-ischemic and inflammatory clinical and MRI biomarkers in MS patients. Here I performed most of the statistical analyses. An abstract with me as second author has been submitted to ECTRIMS[3] 2021. [47]

To conclude this section, I had the chance during this year to facilitate the image post-processing for medical neuroimaging researchers by automating different tasks and by assembling them into pipelines. This resulted in a considerable save of time and gain in accuracy for my medical colleagues in the lab. There is, however, always a way to further improve. My ideas/suggestions for improvement are developed here-below in section 7.3.

7.2 Limitations

A number of factors may limit the results produced in this thesis and more generally the results based on our database. First, and this has already been addressed in section 6.1, the number of missing data is not negligible. To be sure that the imputation method used is reliable, the best way is to compare the machine learning results using different data imputation methods. If the results follow the same trend regardless of the imputation method, we can be confident in these results. Otherwise, the interpretation of the results will be much more difficult. A second limitation is the limited number of patients included in the study, a recurring factor in the field of medical data analysis especially when using advanced-imaging techniques. As a final limitation we can also mention the confidence that one can have in the different ways of measuring our variables. This problem concerns both the variables evaluated by experts (such as the number of PRL, CVS, ...) and those evaluated by programs (e.g. SAMSEG).

7.3 Future steps

As a reminder, this thesis is part of a larger multi-disciplinary project aimed at i) improving the distinction between MS and MS mimicking disorders, ii) identifying different pathological entities on the "MS spectrum," and iii) better predicting conversion to MS in "at risk" patients. The research subjects are almost endless and all deserve to be studied in depth. However, I selected several ways to improve this work or simply come closer to the three aims previously mentioned, and organized them into the 4 main goals this thesis tried to accomplish.

Multi-modal database

[2] http://godzilla.kennedykrieger.org/QSM/
[3] https://www.ectrims.eu/

The first step would be to effectively implement a database management system more suited to our project, according to the requirements developed in section 4.4. In particular, two master students will probably take over this work and hopefully come up with an effective solution during next academic year 2021-2022.

As mentioned in sections 2.2.4 and 7.1, the intralesional and extra-lesional QSM analysis can add significant information to our database and to the overall project (example with [35]). Indeed QSM can give *quantitative* information about tissue integrity in MS, even though its investigation is extremely time consuming.

Another way to have more personalized analyses at the patient level would be to retrieve lesion-level information for each patient. The idea would be to create a "lesion-based" database in order to learn about the relationship between the lesions and the disease.

Insights on MS

The next step to achieve this goal is of course to effectively run the machine learning analysis developed in chapter 6. This will be done in the next months and might end up with an article published in a peer-reviewed scientific journal depending on the results and their reliability.

Automation of MRI feature extraction

An idea to further improve the tools explained in sections 4.2.1 and 4.2.2 would be to add a Graphical User Interface (GUI). Although it might not be the most urgent thing to do, it would simplify significantly the use of these pipelines for researchers with little to no experience with computers or computer science. Additionally, I would like to suggest a naming convention within each MRI center for the different MRI sequences. This would reduce the most limiting source of error of the *Dicom2BIDS* pipeline (section 4.2.1).

Another idea for improvement regards the deep convolutional neural network used to identify PRL called RimNet. Indeed, a pitfall in this approach is that one has to manually separate confluent lesions, i.e. lesions whose origins took place in different areas but merged with other lesions by growing bigger. An idea, already investigated by a master student at the Medical Image Analysis Laboratory, Lausanne, is to use statistical methods [99] to count and separate distinct lesions.

Longitudinal detection of advanced MRI features

The next step to achieve this goal is to develop a new architecture for RimNet that can take into account the longitudinal information. Moreover, we want to try changing the lesion segmentation method to reduce the time splitting confluent lesions manually. Confluent lesions are lesions that were originally distinct but

merged to become one by growing close to each other. The topic of automated lesion splitting is also one of the future steps I would like to explore. I hope I will be able to explore all these intriguing topics during the PhD program I will start in September 2021.

Conclusion

The purpose of this thesis was to progress in four different directions.

The first goal was to establish a multi-modal database that brings together clinical, biological, and especially quite novel data from magnetic resonance imaging. Our study included a set of 135 patients, all with multiple sclerosis, and about 50 variables. To arrive at such a database, we first had to retrieve clinical and biological data from existing databases from different hospitals. Then, we had to extract information from the different images acquired using different imaging protocols. This part was largely the most challenging. To achieve this goal, it was necessary to first organize a database of images in BIDS format. To help me and the other researchers of the laboratory I worked with, I created a tool that allows to transform the raw images coming out of the MRI into more compressed formats and following the BIDS standard.I then developed and used a pipeline ("les_voloc") that uses different image processing tools to calculate the different brain volumes of interest, segment the lesions, calculate their volume and localize them. These programs have also been developed for and used by medical researchers in the lab for several conference abstract publications. In addition, a specification and a solution proposal were elaborated for a database management system whose purpose is to store and manage the numerous data used in this kind of projects.

The second goal of this thesis was to build on the database cited in the previous paragraph with the aim to discover new relationships between the different variables. In addition to simple univariate and bivariate statistics (which still resulted in a clinical abstract regarding microangiopathy in MS patients however), there were two clear long-term subgoals. The first was to provide assistance to neurologists in disease prognosis, i.e., the prediction of disability (EDSS/MSSS/ARMSS) of a patient based on the different factors present in our database. This problem being a regression problem, I proposed several machine learning methods to solve it. The second sub-goal was to find new subtypes of the disease. Indeed, MS patients are currently subdivided according to the way their disease progresses (RRMS, PPMS or SPMS), yet there are probably better ways to subdivide the types of MS. After identifying this problem as one that could be solved by a clustering algorithm, I proposed several of them that could satisfy this objective. In addition, an important point was that despite the collective efforts in filling this database, it could not be filled to 100%, which is of course mandatory for the vast majority of machine learning algorithms. I therefore proposed several data imputation methods to overcome this problem. The next steps (that will probably be done this summer) consist mainly in

the application of the proposed methods.

Third, a goal that this thesis also shares with many other efforts in this field worldwide was to achieve automation of the extraction of features from magnetic resonance imaging in MS patients. The pipelines/tools written to populate the database therefore also served this purpose.

Finally, this thesis was also intended to open up the problem of longitudinal detection of MRI features. Indeed, many studies on MS are based on several years, it is therefore important for programs extracting MRI features such as RimNet and SAMSEG to consider longitudinal information when they process images.

To researchers around the world, multiple sclerosis still holds many mysteries in terms of understanding how it works and how to treat it. During this thesis, some paths among many others have been explored, yet the mystery remains largely intact, unveiling itself little by little as high impact articles are published. In order to better understand and treat this disease, we still need to deepen these paths and also probably explore many more.

Bibliography

[1] C. Lucchinetti et al. “Heterogeneity of multiple sclerosis lesions: implications for the pathogenesis of demyelination”. eng. In: *Annals of Neurology* 47.6 (June 2000), pp. 707–717. ISSN: 0364-5134. DOI: 10.1002/1531-8249(200006)47:6<707::aid-ana3>3.0.co;2-q.

[2] Pietro Maggi et al. “Central vein sign differentiates Multiple Sclerosis from central nervous system inflammatory vasculopathies”. en. In: *Annals of Neurology* 83.2 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.25146, pp. 283–294. ISSN: 1531-8249. DOI: https://doi.org/10.1002/ana.25146. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25146 (visited on 02/02/2021).

[3] Pietro Maggi et al. “The "central vein sign" in patients with diagnostic "red flags" for multiple sclerosis: A prospective multicenter 3T study”. eng. In: *Multiple Sclerosis (Houndmills, Basingstoke, England)* 26.4 (Apr. 2020), pp. 421–432. ISSN: 1477-0970. DOI: 10.1177/1352458519876031.

[4] Tim Sinnecker et al. “Evaluation of the Central Vein Sign as a Diagnostic Imaging Biomarker in Multiple Sclerosis”. eng. In: *JAMA neurology* 76.12 (Dec. 2019), pp. 1446–1456. ISSN: 2168-6157. DOI: 10.1001/jamaneurol.2019.2478.

[5] Pietro Maggi et al. “Paramagnetic Rim Lesions are Specific to Multiple Sclerosis: An International Multicenter 3T MRI Study”. en. In: *Annals of Neurology* 88.5 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.25877, pp. 1034–1042. ISSN: 1531-8249. DOI: https://doi.org/10.1002/ana.25877. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25877 (visited on 02/09/2021).

[6] Martina Absinta et al. “Association of Chronic Active Multiple Sclerosis Lesions With Disability In Vivo”. eng. In: *JAMA neurology* 76.12 (Dec. 2019), pp. 1474–1483. ISSN: 2168-6157. DOI: 10.1001/jamaneurol.2019.2399.

[7] Massimo Filippi et al. “Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines”. In: *Brain* 142.7 (July 2019), pp. 1858–1875. ISSN: 0006-8950. DOI: 10.1093/brain/awz144. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6598631/ (visited on 02/05/2021).

[8] Andrew J. Solomon et al. “MRI evaluation of thalamic volume differentiates MS from common mimics”. In: *Neurology® Neuroimmunology & Neuroinflammation* 4.5 (July 2017). ISSN: 2332-7812. DOI: 10.1212/NXI.0000000000000387. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5515603/ (visited on 05/24/2021).

[9] Pietro Maggi et al. “CVSnet: A machine learning approach for automated central vein sign assessment in multiple sclerosis”. en. In: *NMR in Biomedicine* 33.5 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nbm.4283, e4283. ISSN: 1099-1492. DOI: https://doi.org/10.1002/nbm.4283. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nbm.4283 (visited on 02/12/2021).

[10] Germán Barquero et al. “RimNet: A deep 3D multimodal MRI architecture for paramagnetic rim lesion assessment in multiple sclerosis”. en. In: *NeuroImage: Clinical* 28 (Jan. 2020), p. 102412. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2020.102412. URL: http://www.sciencedirect.com/science/article/pii/S2213158220302497 (visited on 12/13/2020).

[11] Maxence Wynen and al. “Longitudinal automated assesment of paramagnetic rim lesions in multiple sclerosis using RimNet”. In: Online, 2021. URL: https://www.ismrm.org/21m/.

[12] Massimo Filippi et al. “Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines”. In: *Brain* 142.7 (July 2019), pp. 1858–1875. ISSN: 0006-8950. DOI: 10.1093/brain/awz144. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6598631/ (visited on 06/08/2021).

[13] Colm Elliott et al. “Chronic white matter lesion activity predicts clinical progression in primary progressive multiple sclerosis”. eng. In: *Brain: A Journal of Neurology* 142.9 (Sept. 2019), pp. 2787–2799. ISSN: 1460-2156. DOI: 10.1093/brain/awz212.

[14] François Guisset et al. “The central vein sign in multiple sclerosis patients with vascular comorbidities”. eng. In: *Multiple Sclerosis (Houndmills, Basingstoke, England)* 27.7 (June 2021), pp. 1057–1065. ISSN: 1477-0970. DOI: 10.1177/1352458520943785.

[15] Benoit Kastler. *Comprendre l’IRM, Manuel d’auto-apprentissage*. fr. 7e. Elsevier Masson, 2011. ISBN: 978-2-294-71044-5.

[16] Daniel S. Reich, Claudia F. Lucchinetti, and Peter A. Calabresi. “Multiple Sclerosis”. eng. In: *The New England Journal of Medicine* 378.2 (Jan. 2018), pp. 169–180. ISSN: 1533-4406. DOI: 10.1056/NEJMra1401483.

[17] GBD 2015 Neurological Disorders Collaborator Group. “Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015”. eng. In: *The Lancet. Neurology* 16.11 (Nov. 2017), pp. 877–897. ISSN: 1474-4465. DOI: 10.1016/S1474-4422(17)30299-5.

[18] Wallace J. Brownlee et al. “Diagnosis of multiple sclerosis: progress and challenges”. eng. In: *Lancet (London, England)* 389.10076 (Apr. 2017), pp. 1336–1346. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(16)30959-X.

[19] Alastair Compston and Alasdair Coles. “Multiple sclerosis”. eng. In: *Lancet (London, England)* 372.9648 (Oct. 2008), pp. 1502–1517. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(08)61620-7.

[20] Calliope A. Dendrou, Lars Fugger, and Manuel A. Friese. “Immunopathology of multiple sclerosis”. eng. In: *Nature Reviews. Immunology* 15.9 (Sept. 2015), pp. 545–558. ISSN: 1474-1741. DOI: 10.1038/nri3871.

[21] Fred D. Lublin et al. “Defining the clinical course of multiple sclerosis: the 2013 revisions”. eng. In: *Neurology* 83.3 (July 2014), pp. 278–286. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000000560.

[22] B. G. Weinshenker. “Natural history of multiple sclerosis”. eng. In: *Annals of Neurology* 36 Suppl (1994), S6–11. ISSN: 0364-5134. DOI: 10.1002/ana.410360704.

[23] Tobias Derfuss. “Personalized medicine in multiple sclerosis: hope or reality?” In: *BMC Medicine* 10.1 (Oct. 2012), p. 116. ISSN: 1741-7015. DOI: 10.1186/1741-7015-10-116. URL: https://doi.org/10.1186/1741-7015-10-116 (visited on 02/03/2021).

[24] Alan J Thompson et al. “Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria”. en. In: *The Lancet Neurology* 17.2 (Feb. 2018), pp. 162–173. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(17)30470-2. URL: http://www.sciencedirect.com/science/article/pii/S1474442217304702 (visited on 02/02/2021).

[25] Andrew J. Solomon and John R. Corboy. “The tension between early diagnosis and misdiagnosis of multiple sclerosis”. en. In: *Nature Reviews Neurology* 13.9 (Sept. 2017). Number: 9 Publisher: Nature Publishing Group, pp. 567–572. ISSN: 1759-4766. DOI: 10.1038/nrneurol.2017.106. URL: https://www.nature.com/articles/nrneurol.2017.106 (visited on 02/09/2021).

[26] J. F. Kurtzke. “Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS)”. eng. In: *Neurology* 33.11 (Nov. 1983), pp. 1444–1452. ISSN: 0028-3878. DOI: 10.1212/wnl.33.11.1444.

[27] R. H. S. R. Roxburgh et al. “Multiple Sclerosis Severity Score: using disability and disease duration to rate disease severity”. eng. In: *Neurology* 64.7 (Apr. 2005), pp. 1144–1151. ISSN: 1526-632X. DOI: 10.1212/01.WNL.0000156155.19270.F8.

[28] RH Gross et al. “The Multiple Sclerosis Severity Score: fluctuations and prognostic ability in a longitudinal cohort of patients with MS”. In: *Multiple Sclerosis Journal - Experimental, Translational and Clinical* 5.1 (Mar. 2019). ISSN: 2055-2173. DOI: 10.1177/2055217319837254. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6423683/ (visited on 05/24/2021).

[29] Ali Manouchehrinia et al. “Age Related Multiple Sclerosis Severity Score: Disability ranked by age”. In: *Multiple Sclerosis (Houndmills, Basingstoke, England)* 23.14 (Dec. 2017), pp. 1938–1946. ISSN: 1352-4585. DOI: 10.1177/1352458517690618. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5700773/ (visited on 05/24/2021).

[30] Mário João Fartaria. “Partial-volume aware lesion segmentation in Multiple Sclerosis: New image analysis methods for high and ultra-high field MRI”. eng. PhD thesis. Lausanne: Université de Lausanne, 2018.

[31] Bruce Fischl. "FreeSurfer". en. In: *NeuroImage*. 20 YEARS OF fMRI 62.2 (Aug. 2012), pp. 774–781. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2012.01.021. URL: http://www.sciencedirect.com/science/article/pii/S1053811912000389 (visited on 12/16/2020).

[32] E. S. Beck et al. "Improved Visualization of Cortical Lesions in Multiple Sclerosis Using 7T MP2RAGE". eng. In: *AJNR. American journal of neuroradiology* 39.3 (Mar. 2018), pp. 459–466. ISSN: 1936-959X. DOI: 10.3174/ajnr.A5534.

[33] P. Sati et al. "Rapid, high-resolution, whole-brain, susceptibility-based MRI of multiple sclerosis". In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 20.11 (Oct. 2014), pp. 1464–1470. ISSN: 1352-4585. DOI: 10.1177/1352458514525868. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4167170/ (visited on 02/12/2021).

[34] Pietro Maggi et al. "Chronic White Matter Inflammation and Serum Neurofilament Levels in Multiple Sclerosis". en. In: *Neurology* (June 2021). Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Research Article. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.0000000000012326. URL: https://n.neurology.org/content/early/2021/06/04/WNL.0000000000012326 (visited on 06/08/2021).

[35] Reza Rahmanzadeh et al. "Myelin and axon pathology in multiple sclerosis assessed by myelin water and multi-shell diffusion imaging". eng. In: *Brain: A Journal of Neurology* (Mar. 2021). ISSN: 1460-2156. DOI: 10.1093/brain/awab088.

[36] Sabina Luchetti et al. "Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: a retrospective autopsy cohort analysis". eng. In: *Acta Neuropathologica* 135.4 (Apr. 2018), pp. 511–528. ISSN: 1432-0533. DOI: 10.1007/s00401-018-1818-y.

[37] Assunta Dal-Bianco et al. "Slow expansion of multiple sclerosis iron rim lesions: pathology and 7 T magnetic resonance imaging". en. In: *Acta Neuropathologica* 133.1 (Jan. 2017), pp. 25–42. ISSN: 1432-0533. DOI: 10.1007/s00401-016-1636-z. URL: https://doi.org/10.1007/s00401-016-1636-z (visited on 02/09/2021).

[38] Alberto Calvi et al. "In vivo imaging of chronic active lesions in multiple sclerosis". en. In: *Multiple Sclerosis Journal* (Sept. 2020). Publisher: SAGE Publications Ltd STM, p. 1352458520958589. ISSN: 1352-4585. DOI: 10.1177/1352458520958589. URL: https://doi.org/10.1177/1352458520958589 (visited on 10/15/2020).

[39] Ulrike W. Kaunzner et al. "Quantitative susceptibility mapping identifies inflammation in a subset of chronic multiple sclerosis lesions". eng. In: *Brain: A Journal of Neurology* 142.1 (Jan. 2019), pp. 133–145. ISSN: 1460-2156. DOI: 10.1093/brain/awy296.

[40] M. Absinta et al. "Identification of Chronic Active Multiple Sclerosis Lesions on 3T MRI". en. In: *American Journal of Neuroradiology* 39.7 (July 2018). Publisher: American Journal of Neuroradiology Section: Adult Brain, pp. 1233–1238. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A5660. URL: http://www.ajnr.org/content/39/7/1233 (visited on 10/15/2020).

[41] D. M. Harrison et al. “Lesion Heterogeneity on High-Field Susceptibility MRI Is Associated with Multiple Sclerosis Severity”. eng. In: *AJNR. American journal of neuroradiology* 37.8 (Aug. 2016), pp. 1447–1453. ISSN: 1936-959X. DOI: 10.3174/ajnr.A4726.

[42] Alex D. Rae-Grant et al. “Observations on the brain vasculature in multiple sclerosis: A historical perspective”. eng. In: *Multiple Sclerosis and Related Disorders* 3.2 (Mar. 2014), pp. 156–162. ISSN: 2211-0356. DOI: 10.1016/j.msard.2013.08.005.

[43] Michael H. Barnett and John W. Prineas. “Relapsing and remitting multiple sclerosis: pathology of the newly forming lesion”. eng. In: *Annals of Neurology* 55.4 (Apr. 2004), pp. 458–468. ISSN: 0364-5134. DOI: 10.1002/ana.20016.

[44] C. W. Adams. “The onset and progression of the lesion in multiple sclerosis”. eng. In: *Journal of the Neurological Sciences* 25.2 (June 1975), pp. 165–182. ISSN: 0022-510X. DOI: 10.1016/0022-510x(75)90138-0.

[45] Pietro Maggi et al. “The formation of inflammatory demyelinated lesions in cerebral white matter”. eng. In: *Annals of Neurology* 76.4 (Oct. 2014), pp. 594–608. ISSN: 1531-8249. DOI: 10.1002/ana.24242.

[46] S. Suthiphosuwan et al. “The Central Vein Sign in Radiologically Isolated Syndrome”. en. In: *American Journal of Neuroradiology* (Apr. 2019). Publisher: American Journal of Neuroradiology Section: ADULT BRAIN. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A6045. URL: http://www.ajnr.org/content/early/2019/04/18/ajnr.A6045 (visited on 06/08/2021).

[47] Marco Pasi et al. “(In press) Enlarged perivascular spaces on MR views in MS patients are associated to microangiopathy rather than to inflammation”. In: 2021.

[48] J. J. G. Geurts et al. “Consensus recommendations for MS cortical lesion scoring using double inversion recovery MRI”. eng. In: *Neurology* 76.5 (Feb. 2011), pp. 418–424. ISSN: 1526-632X. DOI: 10.1212/WNL.0b013e31820a0cc4.

[49] Francesco La Rosa et al. “Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE”. In: *NeuroImage : Clinical* 27 (June 2020). ISSN: 2213-1582. DOI: 10.1016/j.nicl.2020.102335. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7358270/ (visited on 02/12/2021).

[50] Mário João Fartaria et al. “Automated detection of white matter and cortical lesions in early stages of multiple sclerosis”. eng. In: *Journal of magnetic resonance imaging: JMRI* 43.6 (June 2016), pp. 1445–1454. ISSN: 1522-2586. DOI: 10.1002/jmri.25095.

[51] M. Filippi et al. “Intracortical lesions: relevance for new MRI diagnostic criteria for multiple sclerosis”. eng. In: *Neurology* 75.22 (Nov. 2010), pp. 1988–1994. ISSN: 1526-632X. DOI: 10.1212/WNL.0b013e3181ff96f6.

[52] Paolo Preziosa et al. “Diagnosis of multiple sclerosis: a multicentre study to compare revised McDonald-2010 and Filippi-2010 criteria”. eng. In: *Journal of Neurology, Neurosurgery, and Psychiatry* 89.3 (Mar. 2018), pp. 316–318. ISSN: 1468-330X. DOI: 10.1136/jnnp-2017-315863.

[53] Massimiliano Calabrese et al. “Cortical lesion load associates with progression of disability in multiple sclerosis”. eng. In: *Brain: A Journal of Neurology* 135.Pt 10 (Oct. 2012), pp. 2952–2961. ISSN: 1460-2156. DOI: 10.1093/brain/aws246.

[54] *About DICOM- Overview*. en. URL: https://www.dicomstandard.org/about (visited on 05/11/2021).

[55] *rordenlab/dcm2niix*. original-date: 2014-10-19T17:20:13Z. May 2021. URL: https://github.com/rordenlab/dcm2niix (visited on 05/11/2021).

[56] BIDS-Contributors. “The Brain Imaging Data Structure (BIDS) Specification”. en. In: (Apr. 2021). Publisher: Zenodo Version Number: 1.6.0. DOI: 10.5281/ZENODO.3686061. URL: https://zenodo.org/record/3686061 (visited on 05/11/2021).

[57] *Brain Imaging Data Structure*. en-US. URL: https://bids.neuroimaging.io/index (visited on 05/11/2021).

[58] Barbara Zitová and Jan Flusser. “Image registration methods: a survey”. en. In: *Image and Vision Computing* 21.11 (Oct. 2003), pp. 977–1000. ISSN: 02628856. DOI: 10.1016/S0262-8856(03)00137-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0262885603001379 (visited on 05/11/2021).

[59] Baudouin Senneville et al. “EVolution: an Edge-based Variational method for non-rigid multi-modal image registration”. In: *Physics in Medicine and Biology* 61 (Sept. 2016). DOI: 10.1088/0031-9155/61/20/7377.

[60] B. B. Avants et al. “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain”. eng. In: *Medical Image Analysis* 12.1 (Feb. 2008), pp. 26–41. ISSN: 1361-8423. DOI: 10.1016/j.media.2007.06.004.

[61] Abdullah Nazib et al. “Performance of Image Registration Tools on High-Resolution 3D Brain Images”. In: *arXiv:1807.04917 [cs]* (July 2018). arXiv: 1807.04917. URL: http://arxiv.org/abs/1807.04917 (visited on 05/10/2021).

[62] Ivana Despotović, Bart Goossens, and Wilfried Philips. “MRI segmentation of the human brain: challenges, methods, and applications”. eng. In: *Computational and Mathematical Methods in Medicine* 2015 (2015), p. 450341. ISSN: 1748-6718. DOI: 10.1155/2015/450341.

[63] Stefano Cerri et al. “A Contrast-Adaptive Method for Simultaneous Whole-Brain and Lesion Segmentation in Multiple Sclerosis”. In: *NeuroImage* 225 (Jan. 2021). arXiv: 2005.05135, p. 117471. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2020.117471. URL: http://arxiv.org/abs/2005.05135 (visited on 12/13/2020).

[64] Donatas Sederevičius et al. “Reliability and sensitivity of two whole-brain segmentation approaches included in FreeSurfer - ASEG and SAMSEG”. eng. In: *NeuroImage* (Apr. 2021), p. 118113. ISSN: 1095-9572. DOI: 10.1016/j.neuroimage.2021.118113.

[65] *Pydicom /*. URL: https://pydicom.github.io/ (visited on 05/11/2021).

[66] *Neuroimaging in Python — NiBabel 3.2.0 documentation*. URL: https://nipy.org/nibabel/ (visited on 05/11/2021).

[67] *SciPy.org — SciPy.org*. URL: https://www.scipy.org/ (visited on 05/11/2021).

[68] *motioncor - Free Surfer Wiki*. URL: https://surfer.nmr.mgh.harvard.edu/fswiki/motioncor (visited on 05/11/2021).

[69] *pandas - Python Data Analysis Library*. URL: https://pandas.pydata.org/ (visited on 05/24/2021).

[70] *NumPy*. URL: https://numpy.org/ (visited on 05/24/2021).

[71] *scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation*. URL: https://scikit-learn.org/stable/ (visited on 05/24/2021).

[72] *SciPy.org — SciPy.org*. URL: https://www.scipy.org/ (visited on 05/24/2021).

[73] *seaborn: statistical data visualization — seaborn 0.11.1 documentation*. URL: https://seaborn.pydata.org/ (visited on 05/24/2021).

[74] Ralph B. D’Agostino. “An Omnibus Test of Normality for Moderate and Large Size Samples”. In: *Biometrika* 58.2 (1971). Publisher: [Oxford University Press, Biometrika Trust], pp. 341–348. ISSN: 0006-3444. DOI: 10.2307/2334522. URL: https://www.jstor.org/stable/2334522 (visited on 05/26/2021).

[75] Ralph D’Agostino and E. S. Pearson. “Tests for Departure from Normality. Empirical Results for the Distributions of b2 and $\sqrt{}$ b1”. In: *Biometrika* 60.3 (1973). Publisher: [Oxford University Press, Biometrika Trust], pp. 613–622. ISSN: 0006-3444. DOI: 10.2307/2335012. URL: https://www.jstor.org/stable/2335012 (visited on 05/26/2021).

[76] S. S. Shapiro and M. B. Wilk. “An Analysis of Variance Test for Normality (Complete Samples)”. In: *Biometrika* 52.3/4 (1965). Publisher: [Oxford University Press, Biometrika Trust], pp. 591–611. ISSN: 0006-3444. DOI: 10.2307/2333709. URL: https://www.jstor.org/stable/2333709 (visited on 05/26/2021).

[77] Sandra González-Villà et al. “Evaluating the effect of multiple sclerosis lesions on automatic brain structure segmentation”. en. In: *NeuroImage: Clinical* 15 (Jan. 2017), pp. 228–238. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2017.05.003. URL: https://www.sciencedirect.com/science/article/pii/S2213158217301080 (visited on 06/10/2021).

[78] Sergi Valverde, Arnau Oliver, and Xavier Lladó. “A white matter lesion-filling approach to improve brain tissue volume measurements”. en. In: *NeuroImage: Clinical* 6 (Jan. 2014), pp. 86–92. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2014.08.016. URL: https://www.sciencedirect.com/science/article/pii/S2213158214001259 (visited on 06/10/2021).

[79] M. Battaglini, M. Jenkinson, and N De Stefano. “Evaluating and reducing the impact of white matter lesions on brain volume measurements”. English. In: *Human Brain Mapping* (Sept. 2012), pp. 2062–2071. ISSN: 10659471.

[80] Stephen M. Smith. “Fast robust automated brain extraction”. eng. In: *Human Brain Mapping* 17.3 (Nov. 2002), pp. 143–155. ISSN: 1065-9471. DOI: 10.1002/hbm.10062.

[81] Stéfan van der Walt et al. “scikit-image: image processing in Python”. en. In: *PeerJ* 2 (June 2014). Publisher: PeerJ Inc., e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://peerj.com/articles/453 (visited on 12/13/2020).

[82] Ramez Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems*. en-US. 7th Edition. Pearson, 2016. ISBN: 978-0-13-397077-7.

[83] M. Matsumae et al. “Age-related changes in intracranial compartment volumes in normal adults assessed by magnetic resonance imaging”. eng. In: *Journal of Neurosurgery* 84.6 (June 1996), pp. 982–991. ISSN: 0022-3085. DOI: 10.3171/jns.1996.84.6.0982.

[84] Danai Dima et al. “Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3-90 years”. eng. In: *Human Brain Mapping* (Feb. 2021). ISSN: 1097-0193. DOI: 10.1002/hbm.25320.

[85] Alastair Compston and Alasdair Coles. “Multiple sclerosis”. eng. In: *Lancet (London, England)* 359.9313 (Apr. 2002), pp. 1221–1231. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(02)08220-X.

[86] F Pérez-Miralles et al. “Clinical impact of early brain atrophy in clinically isolated syndromes”. en. In: *Multiple Sclerosis Journal* 19.14 (Dec. 2013). Publisher: SAGE Publications Ltd STM, pp. 1878–1886. ISSN: 1352-4585. DOI: 10.1177/1352458513488231. URL: https://doi.org/10.1177/1352458513488231 (visited on 06/10/2021).

[87] Maria Pia Sormani, Douglas L. Arnold, and Nicola De Stefano. “Treatment effect on brain atrophy correlates with treatment effect on disability in multiple sclerosis”. en. In: *Annals of Neurology* 75.1 (2014). _eprint: https://onlinelibrary.wiley.com/doi/p
pp. 43–49. ISSN: 1531-8249. DOI: 10.1002/ana.24018. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.24018 (visited on 06/10/2021).

[88] Amina Sellimi et al. “Chronic active MRI lesions in progressive multiple sclerosis patients undergoing ocrelizumab treatment”. In: Online, 2021.

[89] John W. Graham. “Missing data analysis: making it work in the real world”. eng. In: *Annual Review of Psychology* 60 (2009), pp. 549–576. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.58.110405.085530.

[90] Joseph L. Schafer and John W. Graham. “Missing data: our view of the state of the art”. eng. In: *Psychological Methods* 7.2 (June 2002), pp. 147–177. ISSN: 1082-989X.

[91] Abraham Bookstein, Vladimir Kulyukin, and Timo Raita. “Generalized Hamming Distance”. In: *Information Retrieval* 5 (Oct. 2002). DOI: 10.1023/A:1020499411651.

[92] Tanzila Islam et al. “A Deep Learning Method to Impute Missing Values and Compress Genome-wide Polymorphism Data in Rice:” en. In: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*. Online Streaming, — Select a Country —: SCITEPRESS - Science and Technology Publications, 2021, pp. 101–109. ISBN: 978-989-758-490-9. DOI: 10.5220/0010233901010109. URL: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010233901010109 (visited on 05/28/2021).

[93] Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations in R”. en. In: *Journal of Statistical Software* 45.1 (Dec. 2011). Number: 1, pp. 1–67. ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03. URL: https://www.jstatsoft.org/index.php/jss/article/view/v045i03 (visited on 05/28/2021).

[94] Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. “SICE: an improved missing data imputation technique”. In: *Journal of Big Data* 7.1 (2020). ISSN: 2196-1115. DOI: 10.1186/s40537-020-00313-w. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7291187/ (visited on 05/28/2021).

[95] Adrian Tousignant et al. “Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data”. en. In: (), p. 10.

[96] Mauro F. Pinto et al. “Prediction of disease progression and outcomes in multiple sclerosis with machine learning”. en. In: *Scientific Reports* 10.1 (Dec. 2020). Number: 1 Publisher: Nature Publishing Group, p. 21038. ISSN: 2045-2322. DOI: 10.1038/s41598-020-78212-6. URL: https://www.nature.com/articles/s41598-020-78212-6 (visited on 05/28/2021).

[97] Ruggiero Seccia et al. “Machine Learning Use for Prognostic Purposes in Multiple Sclerosis”. eng. In: *Life (Basel, Switzerland)* 11.2 (Feb. 2021). ISSN: 2075-1729. DOI: 10.3390/life11020122.

[98] Arman Eshaghi et al. “Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data”. en. In: *Nature Communications* 12.1 (Apr. 2021). Number: 1 Publisher: Nature Publishing Group, p. 2078. ISSN: 2041-1723. DOI: 10.1038/s41467-021-22265-2. URL: https://www.nature.com/articles/s41467-021-22265-2 (visited on 05/28/2021).

[99] J. D. Dworkin et al. “An Automated Statistical Technique for Counting Distinct Multiple Sclerosis Lesions”. en. In: *American Journal of Neuroradiology* (Feb. 2018). Publisher: American Journal of Neuroradiology Section: ADULT BRAIN. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A5556. URL: http://www.ajnr.org/content/early/2018/02/22/ajnr.A5556 (visited on 06/01/2021).

Appendices

Appendix A

Memory space used for one MRI session in MS

Subject ID	DICOM directory size (MB)	Space taken in BIDS directory (MB)
14	1900	670,897647
16	1290	466,413338
17	1040	452,240655
18	708	465,24964
19	1200	849,818816
21	2240	403,249119
22	683	292,597724
23	351	719,432576
25	1410	463,017946
27	2330	615,327264
34	1130	859,848906
35	1060	838,574821
36	753	645,460553
36	403	645,460553
39	93	578,833348
41	993	789,836865
43	1080	801,61919
45	747	857,27528
47	912	882,429673
48	927	844,949158
49	996	649,39392
54	993	735,211614
56	1040	637,88463
57	1000	678,91588
59	556	768,077481
60	761	680,541116
61	866	666,843875
62	761	680,870925
64	1000	651,570152
66	1000	687,191822
67	970	665,778798
68	972	691,244313
69	980	693,550208
70	1070	686,449399
71	704	667,826115
72	1000	685,61946
73	888	719,287386
74	1060	699,607451
76	1030	691,417237
120	1230	873,427891
130	535	455,793423

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**