Patch-based MRI Analysis: From voxel to knowledge

Pierrick Coupé

université de BORDEAUX

Habilitation à diriger des recherches

Université de Bordeaux

Presented

by Pierrick COUPE

Patch-based MRI Analysis:

From voxel to knowledge

Defended on May 29th, 2018

Jury members:

M.	AYACHE	Nicholas	DR INRIA, INRIA Sophia Antipolis, FR.	Président
Mme.	BLOCH	Isabelle	PU, Telecom ParisTech, FR.	Rapporteuse
M.	DOJAT	Michel	DR INSERM, Univ. Joseph Fourier, FR.	Rapporteur
M.	HAMMERS	Alexander	PU, King's College, UK.	Rapporteur
M.	COLLINS	Louis	PU, McGill University, CA.	Examinateur
M.	DOMENGER	Jean-Philippe	PU, Université de Bordeaux, FR.	Examinateur

Titre : Analyse par patchs des IRM : du voxel à la connaissance

Résumé : Cette Habilitation à Diriger des Recherches présente les travaux que j'ai menés sur l'analyse quantitative des IRM, le diagnostic assisté par ordinateur et le monitoring cérébral. Mes contributions dans ces domaines sont détaillées dans 3 chapitres. Dans le premier chapitre, j'introduis le principe de notre méthode de segmentation par patchs ainsi que plusieurs extensions. Ensuite, j'analyse les résultats obtenus par nos méthodes pour plusieurs applications. Dans le deuxième chapitre, je montre comment nous avons étendu notre méthode de segmentation par patchs à la détection de pathologies. Les performances de cette méthode sont évaluées pour la détection et la prédiction de la maladie d'Alzheimer. Dans le troisième chapitre, je décris les outils que nous avons développés dans le but d'effectuer un monitoring cérébral. Tout d'abord, la chaine de traitement proposée pour l'analyse quantitative du cerveau est détaillée. Ensuite, les modèles standard proposés afin de déterminer si le volume d'une structure cérébrale est normal ou non sont présentés. De plus, les nouvelles connaissances médicales et neuroscientifiques sur le développement et le vieillissement du cerveau produites lors de leur élaboration sont analysées. Enfin, je décris la plate-forme volBrain en libre accès que nous avons développée. Pour conclure ce manuscrit, je discute des limites et des perspectives de ma recherche sur l'analyse des IRM par patchs.

Mots clés : Imagerie médicale, Neuroimagerie, Segmentation, Aide au diagnostic, IRM, Cerveau, Maladie d'Alzheimer.

Title: Patch-based MRI Analysis: From voxel to knowledge

Abstract: This Habilitation thesis presents the work that I have done on quantitative MR analysis, computer-aided diagnosis and brain monitoring. These topics are detailed in 3 chapters. In the first chapter, I introduce the principle of our patch-based segmentation method and their extensions. Afterwards, the results obtained by our patch-based segmentation method are analyzed for several applications. In the second chapter, I show how we have extended our patch-based segmentation framework to patch-based grading of brain structures. Then, the performance of our patch-based grading method to achieve Alzheimer's disease diagnosis and prognosis is evaluated. In the third chapter, I describe the tools that we developed to perform brain monitoring. First, the pipeline proposed to perform quantitative brain analysis are detailed. Second, the construction of the standard models is presented. Moreover, new medical and neuroscientific knowledge on the development and the aging of the brain produced during their estimation are discussed. Finally, the developed open access volBrain platform is described. To conclude this manuscript, I discuss the limitations and the perspectives of my research about patch-based MRI analysis.

Keywords: Medical imaging, Neuroimaging, Patch-based segmentation, Patch-based grading, MRI, Brain, Computer-aided diagnosis, Brain monitoring

Table of Contents

List of Abbreviations

AD	Alzheimer's Disease
ADAS	Alzheimer’s Disease Assessment Scale
ADNI	Alzheimer’s Disease Neuroimaging Initiative
ANN	Approximate Nearest Neighbor
BEaST	Brain Extraction based on nonlocal Segmentation Technique
CERES	Cerebellum Lobules Segmentation
CSF	Cerebro Spinal Fluid
CN	Cognitive Normal subject
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CA	Cornu Ammonis
CTH	Cortical THickness measurements
DL	Deep Learning
DLB	dementia with Lewy bodies
DG	Dentate Gyrus
DTI	Diffusion Tensor Imaging
ENTCOR	Entorhinal Cortex
FLAIR	Fluid-Attenuated Inversion Recovery
FTD	frontotemporal dementia
FCN	Fully Convolutional Network
GM	Gray Matter
HIPS	Hippocampal Subfields Segmentation
HIPP	Hippocampus
ICBM	International Consortium for Brain Mapping
ICC	IntraCranial Cavity
LNOCV	Leave-N-Out Cross Validation
LOOCV	Leave-One-Out Cross Validation
LDA	Linear Discriminant Analysis
MRI	Magnetic Resonance Images
MD	Mean Diffusivity
MTL	Medial Temporal Lobe
MMSE	Mini Mental State Evaluation
MPBG	Multimodal Patch-based Grading
MS	Multiple Sclerosis
NIHPD	NIH Pediatric Database
NICE	Non-local Intracranial Cavity Extraction
OPM	Optimized PatchMatch
OPAL	Optimized PAtchMatch for Label fusion
PVC	Partial Volume Coefficients
PBG	Patch-Based Grading

PBS	Patch-Based Segmentation
PM	PatchMatch
MCI	patient Mild Cognitive Impairment
eMCI	patient with early Mild Cognitive Impairment
lMCI	patient with late Mild Cognitive Impairment
pMCI	patient with progressive Mild Cognitive Impairment
sMCI	patient with stable Mild Cognitive Impairment
PVD	Percent Volume Difference
PET	Positron Emission Tomography
RRMS	Relapsing Remitting Multiple Sclerosis
SaaS	Software-as-a-Service
SLR	Sparse Logistic Regression
SL	Stratum lacunosum
SM	Stratum molecuare
SR	Stratum radiatum
SSIM	Structural SIMilarity
SSD	Sum of the Squared Difference
SVM	Support Vector Machine
T1w	T1-weighted
T2w	T2-weighted
TBM	Tensor-based Motphometry
TIV	Total Intracranial Volume
VaD	Vascular Dementia
VBM	Voxel-based Morphometry

General Introduction

Over the past 15 years, I have been interested in problems related to medical images analysis involving image processing, applied mathematics, computer science, neuroimaging, neuroanatomy, neuroscience and clinical research. I tried to develop new tools to assist and help clinicians in their task. Moreover, I attempted to propose robust and accurate methods in order to have a real impact in society. This forced me to make significant efforts to develop, validate, distribute, disseminate and valorize all the methods developed. My research on medical image analysis covers different fields, from preprocessing to automatic prognosis of diseases. My work mainly focuses on four topics: image enhancement, quantitative MR analysis, computer-aided diagnosis and brain monitoring. My first major contribution to the field of medical image analysis was done during my Ph.D. with the adaptation of the nonlocal means denoising filter (Buades et al. 2005) to 3D MRI denoising (Coupe et al. 2006, Coupe et al. 2008b). As shown in this manuscript, most of my following contributions have been based on this first work since exemplar-based methods have inspired me for a long time. Although, I have pursued my work on image enhancement (Wiest-Daessle et al. 2007, Coupe et al. 2008a, Wiest-Daessle et al. 2008, Coupe et al. 2009a, Coupe et al. 2009b, Coupe et al. 2010b, Manjon et al. 2010a, Manjon et al. 2010b, Manjon et al. 2010c, Coupe et al. 2012c, Manjon et al. 2012, Coupe et al. 2013, Manjon et al. 2013, Guizard et al. 2015b, Manjon et al. 2015, St-Jean et al. 2016), I decided not to include this topic in this habilitation thesis and to focus only on my recent research achievements. Therefore, this document will present the works carried out over the past 8 years dedicated to quantitative MR analysis, computer-aided diagnosis and brain monitoring. These selected topics will be detailed in 3 chapters.

In the first chapter, I will introduce the principle of the patch-based segmentation (PBS) method that we introduced in (Coupe et al. 2010a, Coupe et al. 2011). This nonlocal patch-based approach uses expert's manual segmentations as priors to segment a new case. I will highlight the link between the proposed PBS and the original nonlocal means denoising filter. In addition, I will explain the paradigm shift between the two frameworks, from nonlocal self-similarity to nonlocal inter-subject similarity. Indeed, in (Coupe et al. 2010a, Coupe et al. 2011), we were the first to apply nonlocal means strategy for segmentation by searching similar patches between the subject to be segmented and a library of training subjects. Then, I will present different extensions that we proposed such as multiscale framework for brain extraction (Eskildsen et al. 2012, Manjon et al. 2014), fast patch search strategy for near real time segmentation (Ta et al. 2014, Giraud et al. 2016), multimodal extension for lesion detection (Guizard et al. 2015a) or combination of fast PBS with multiple nonlinear registrations for cerebellum lobules and hippocampal subfields segmentation. Afterwards, I will show the results obtained by our PBS method and the proposed improvements for different applications. For each considered application, our PBS framework will be compared with *state-of-the-art* methods. I will show that our PBS is now considered as a *state-of-the-art* method for anatomical structure segmentation and is studied by several groups in the world (Bai et al. 2013, Wang and Yushkevich 2013, Wang et al. 2013b, Wolz et al. 2013, Wang et al. 2014b, Wu et al. 2014, Tong et al. 2015, Wu et al. 2015b). Finally, I will discuss the limitations and perspectives of our PBS.

In the second chapter, I will show how we have extended our patch-based segmentation framework to patch-based grading (PBG) of anatomical structures (Coupe et al. 2012a, Coupe et al. 2012b). The grading of the structure under consideration is achieved by estimating the nonlocal similarity of the subject to different training populations. Our PBG method estimates at each voxel a score/grade reflecting the degree/severity of the pathology and thus enables to perform computer-aided diagnosis. I will detail this second paradigm shift from the nonlocal inter-subject similarity used for segmentation to the nonlocal similarity to a population used for grading/scoring. Indeed, when we search for similar patches between the subject to be classified and different training populations, we aim to infer the class of the testing subject by finding the closest training population in the nonlocal sense. After a general principle description, I will present different extensions that we developed such as ensemble-based grading (Komlagan et al. 2014), multimodal grading (Hett et al. 2016, Hett et al. 2018) or multi-feature strategy (Hett et al. 2017). Then, I will present the application of our PBG method to Alzheimer's disease (AD) diagnosis and prognosis. Different experiments will be detailed to validate the developed computer-aided diagnosis tools. Finally, I will present the prognosis performance of our method at presymptomatic phase using a long follow-up dataset (Coupé et al. 2015). This last experiment will enable to study anatomical alterations occurring more than seven years before conversion to AD. While usual volumetric approaches fail to detect anatomical modifications at this presymptomatic stage, I will show that our PBG is an efficient tool to detect such early signs of AD.

In the third chapter, I will present the tools that we developed to perform brain monitoring. Brain monitoring is possible only when two elements are available – a tool to measure the studied parameter and a standard model to know when the obtained measurement is normal or not. First, I will describe the volBrain pipeline that we proposed to perform quantitative brain analysis (Manjon and Coupe 2016). This pipeline provides the volume of brain structures at different scales from tissues to structures. Moreover, the volBrain pipeline includes most of my previous works from denoising (Manjon et al. 2010c) to segmentation (Coupe et al. 2011, Manjon et al. 2014, Romero et al. 2015). Therefore, this pipeline is a good illustration of the complementary of my research topics. Second, I will present the standard models for anatomical brain structures that we estimated to determine when measurements are normal or not (Coupe et al. 2017). Thanks to the new paradigm of Big Data sharing in neuroimaging, we have developed standard models across the entire lifespan based on a massive number of freely available MRI. For the first time, we proposed a unified analysis of brain development and aging from few months of life to advanced ages. In addition, I will present pathological model that we have inferred across the entire lifespan using a similar framework (Coupé et al. 2018). These recent works show the path I have followed over the past years from voxelwise analysis to the production of new knowledge. Finally, I will present the proposed web-platform that integrates the volBrain pipeline and the developed lifespan models. This final part highlights the important work that I dedicated to translational research by proposing innovative ways to reduce the time between methodological developments and their worldwide use.

To conclude this manuscript, I will discuss the limitations and the perspectives of my research about patch-based MRI analysis. Moreover, I will identify the next challenges related to brain MRI analysis and I will question the next evolutions of my work. Finally, I will show how the proposed solutions could be useful for upcoming challenge related to Big Data and Artificial Intelligence.

Chapter 1 Quantitative MRI Analysis

Abstract: In this first chapter, we will introduce the principle of our patch-based segmentation (PBS) method. Inspired by the nonlocal means denoising filter, our nonlocal patch-based approach proposes to use expert's manual segmentations as priors in the context of anatomical structure segmentation. Afterwards, we will present different improvements such as multiscale framework, fast patch search strategy, multi-feature framework or multimodal extension. These improvements address a wide range of problems from brain extraction to multimodal lesion detection. We will show that the proposed improvements drastically reduce the computational time and significantly improve segmentation accuracy of PBS. For each considered application, our PBS framework will be compared with state-of-the-art methods. Finally, we will demonstrate the competitive performances of the proposed PBS for different clinical applications.

1.1 Introduction

Magnetic resonance (MR) imaging plays a crucial role in the detection of pathology, the study of brain organization, and clinical research. Every day, a vast amount of data is produced in clinical settings, preventing the use of manual approaches for data analysis. Consequently, the development of accurate, robust, and reliable segmentation techniques for the automatic extraction of anatomical structures is becoming an important challenge in quantitative MR analysis. In contrast to brain tissue classification where the intensity of the MR signal can be used to segment different tissue types, anatomical segmentation usually requires information derived from the manual segmentations done by experts (i.e., expert priors), since anatomical structures can be composed of several tissue types and distinct anatomical structures can have the same MR signal properties.

To overcome this difficulty, several automatic methods of segmentation have been proposed, such as deformable models or region growing (Ghanei et al. 1998, Shen et al. 2002, Chupin et al. 2007), appearance-based models (Duchesne et al. 2002, Hu and Collins 2007) and atlas/template-warping techniques (Collins et al. 1995, Fischl et al. 2002, Rohlfing et al. 2004, Zhou and Rajapakse 2005, Heckemann et al. 2006, Hammers et al. 2007, Barnes et al. 2008, Gousias et al. 2008, Aljabar et al. 2009, Heckemann et al. 2010).

Among these methods, the atlas-based method has been very successful over the last decades (Collins et al. 1995, Hammers et al. 2002). Indeed, this method allows easy integration of expert priors through the use of a manually segmented atlas. This atlas (i.e., average of many MR images with manual expert-based segmentation) is nonlinearly registered to the subject to be segmented. Then, the estimated deformation field is applied to the manual segmentation. As a result, the manual segmentation is deformed to fit the subject's anatomy (see Figure 1). Thanks to its robustness and simplicity, this method is still used more than 20y after its publication.

More recently, template-warping techniques that use a library of templates (i.e., MR images with manual expert-based segmentation) in place of a single atlas have been the subject of intensive investigation for their high accuracy in segmenting anatomical structures. Barnes et al. (2008) proposed to register the most similar template from a library of prelabeled subjects to segment the hippocampus (HIPP) (see Figure 2). However, the use of only one template may result in a biased segmentation.

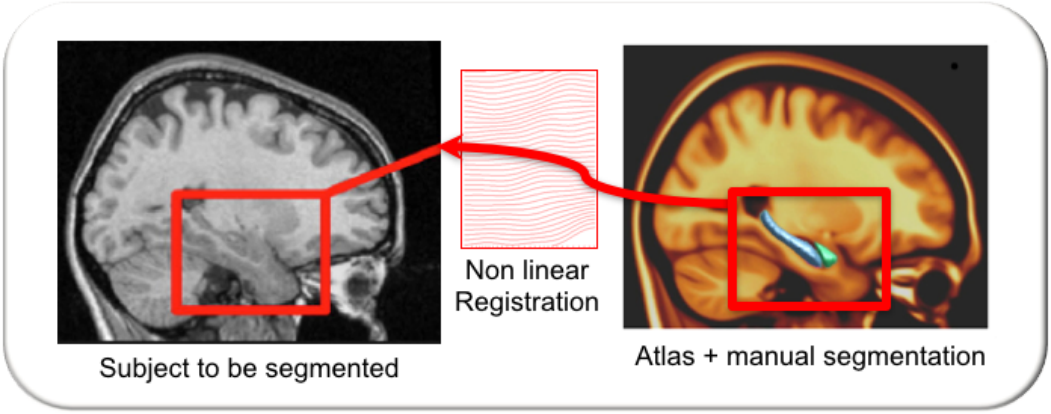

Figure 1: Principle of Atlas-based segmentation. First, the atlas is nonlinearly registered to the subject to be segmented. Then, the estimated deformation field is applied to the manual segmentation of the atlas. Adapted from (Collins and Pruessner 2010).

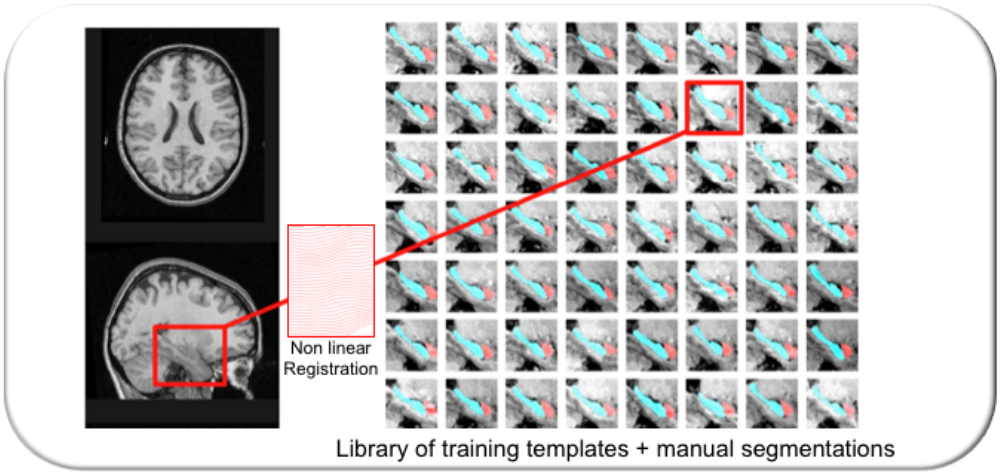

Figure 2: Principle of best template segmentation. First, the most similar template is nonlinearly registered to the subject to be segmented. Then, the estimated deformation field is applied to the manual segmentation. Adapted from (Collins and Pruessner 2010).

To avoid this problem, it is possible to use several similar templates. Pioneering work dedicated to brain segmentation based on multi-template framework has been proposed in (Hammers et al. 2003, Heckemann et al. 2006). In such a method, a set of segmentations is obtained by registering several templates. Therefore, this requires a label fusion strategy to efficiently merge the information derived from the selected templates (see Figure 3). Many strategies have been proposed in the past to achieve this label fusion step (Heckemann et al. 2006, Hammers et al. 2007, Gousias et al. 2008, Aljabar et al. 2009, Collins and Pruessner 2010, Heckemann et al. 2010, Lotjonen et al. 2010). Multi-template segmentation obtained very good performance and is the current *state-of-the-art* approach. However, two main assumptions are made in template-warping techniques. First, constraints on the shapes of structures are used implicitly because of the one-to-one correspondence between the voxels of the image to be segmented and those of the warped templates. This restriction presents the advantage of forcing the resulting segmentation to have a similar shape to those of expert-labeled structures in the template library. However, according to the regularization used during registration, some details can be lost and local high variability cannot be captured. Second, label fusion techniques usually assign the same weight to all the training templates. This approach is sensitive to registration error, since it does not consider the relevance of each template.

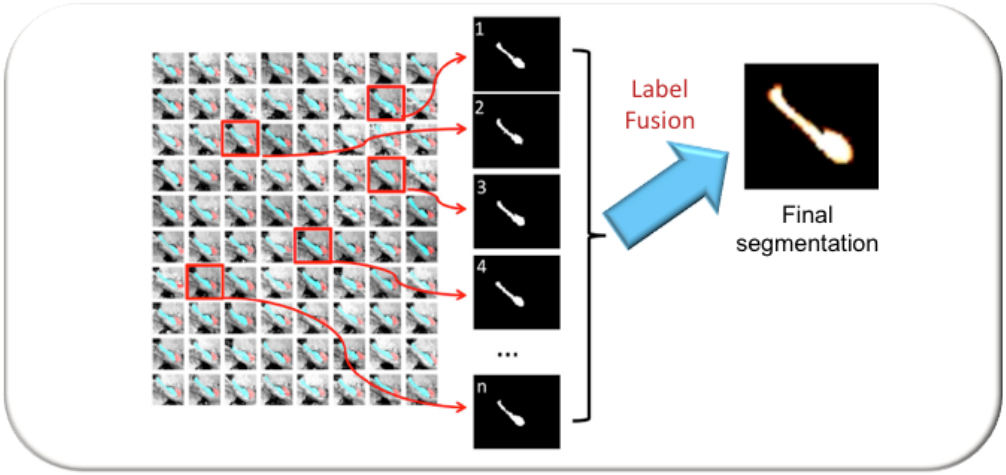

Figure 3: Principle of multi-template segmentation. First, the n most similar templates are nonlinearly registered to the subject to be segmented. Then, the estimated deformation fields are applied to the corresponding manual segmentations of the selected templates. Courtesy of Pr. Collins.

In (Coupe et al. 2010a, Coupe et al. 2011), we proposed to overcome these issues by introducing a novel patch-based scheme with a weighted label fusion, where the weight of each sample is only driven by the similarity of intensity between patches (i.e., small sub-volumes of the image defined as three-dimensional cubes). In the proposed method, voxels with similar surrounding neighborhoods are considered to belong to the same structure and thus are used to estimate the final label. At those time, patch-based methods were the focus of attention of the computer vision community in various domains such as texture synthesis (Efros and Freeman 2001), in-painting (Criminisi et al. 2004), restoration (Buades et al. 2005), and single-frame super resolution (Protter et al. 2009). In each of these domains, patch-based methods have been the subject of intensive investigation because they exhibit very high performance despite their simplicity. Inspired by the nonlocal means denoising filter (Buades et al. 2005), we proposed a nonlocal patch-based approach using expert manual segmentations as priors in the context of anatomical segmentation (Coupe et al. 2011). The nonlocal means filter has two interesting properties that can be exploited to improve segmentation. First, the natural redundancy of information contained in the image can be used to drastically increase the numbers of samples considered during estimation. Second, the local intensity context (i.e., patch) can be used to produce a robust comparison of samples.

In this chapter, we will introduce the principle of our patch-based segmentation method. Afterwards, we will present different improvements such as multiscale framework (Eskildsen et al. 2012, Manjon et al. 2014), fast patch search strategy, multi-feature framework (Giraud et al. 2016) or multimodal extension (Guizard et al. 2015a). Finally, we will discuss the main results for different clinical applications where we applied PBS such as Alzheimer's disease (AD) (Coupe et al. 2012a, Coupe et al. 2012b, Tong et al. 2013, Bron et al. 2015) or Multiple sclerosis (MS) (Guizard et al. 2015a, Planche et al. 2017).

1.2 Methods

1.2.1 The Nonlocal Means Estimator

The nonlocal means filter was first introduced by Buades *et al.* (2005) for the purpose of image denoising. In nonlocal means-based approaches (Buades et al. 2005, Coupe et al. 2008b), the patch $P(x_i)$ surrounding the voxel x_i under study is compared with all the patches $P(x_j)$ of the image Ω whatever their spatial distance to $P(x_i)$ (i.e., this is the meaning of the term "nonlocal"). According to the patch similarity between $P(x_i)$ and $P(x_j)$, estimated by the sum of squared differences (SSD) measure, each patch receives a weight $w(x_i, x_j)$:

$$w(x_i, x_j) = e^{-\frac{\|P(x_i)-P(x_j)\|_2^2}{h^2}} \quad \text{(Eq. 1)}$$

where $\|.\|_2$ is the L2-norm computed between each intensity of the elements of the patches $P(x_i)$ and $P(x_j)$, and h^2 is the smoothing parameter of the weighting function. Finally, all the intensities $u(x_j)$ of the central voxels of the patches $P(x_j)$ are aggregated through a weighted average using the weights $w(x_i, x_j)$. In this way, the denoised intensity $\hat{u}(x_i)$ of the voxel x_i can be efficiently estimated by:

$$\hat{u}(x_i) = \frac{\sum_{j\in\Omega} w(x_i, x_j)u(x_j)}{\sum_{j\in\Omega} w(x_i, x_j)} \quad \text{(Eq. 2)}$$

Despite its simplicity, the nonlocal means filter has been demonstrated to have excellent denoising performance. This filter was one of the most studied denoising filters and many improvements have been proposed since its introduction (see (Buades et al. 2010) for a review of these improvements). The efficiency of the nonlocal means filter relies on two intuitive aspects, the pattern redundancy present in an image (i.e., its self-similarity) and the robust detection of samples derived from the same population by using local context (i.e., patch-based comparison).

First, to improve the accuracy of an estimator, it is possible to reduce the committed error by increasing the number of involved samples. By using an infinite number of samples derived from the same population, the error theoretically converges to zero. To drastically increase the number of samples used, the nonlocal means filter takes advantage of the redundancy of information by using all the similar voxels present over the entire image.

Second, to ensure that the used samples are derived from the same population, the surrounding neighbor of a voxel can be used to robustly detect similar realizations of the same process. In the nonlocal means approach, this task is achieved by patch-based comparison using SSD. Two voxels with similar surrounding patches are considered as similar and to belong to the same population. More precisely, the nonlocal means filter performs patch comparison to estimate the degree of the similarity between two voxels. This way, each involved sample has a weight (see Eq. 1) reflecting its relevance.

Finally, a simple weighted average (see Eq. 2) is used to aggregate the samples according to their relevance. This way, the resulting estimator embodies the two interesting qualities described above: to build on a large number of samples and to ensure that the involved samples are derived from the same population.

1.2.2 From Denoising to Segmentation

In (Coupe et al. 2010a, Coupe et al. 2011), we were the first to introduce the nonlocal means estimator in the context of segmentation by averaging labels instead of intensities. By using a training library of N subjects, whose segmentations of structures are known (see Figure 4), the weighted label fusion is estimated as follows:

$$v(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j}) l(x_{s,j})}{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j})} \quad \text{(Eq. 3)}$$

where $l(x_{s,j})$ is the label (i.e., 0 for background and 1 for structure) given by the expert to the voxel $x_{s,j}$ at location j in the training subject s. It has been shown that the nonlocal means estimator $v(x_i)$ provides a robust estimation of the expected label at x_i. With a label set of 0 and 1, voxels with value $v(x_i) \geq 0.5$ are considered as belonging to the structure and the remaining voxels as background.

As in multi-template segmentation methods, the proposed patch-based method uses expert manual segmentations as priors to achieve the segmentation of anatomical structures. However, our method has two main differences compared with template-warping methods: the scale of the considered objects and the label fusion scheme.

First, while multi-template methods work at the level of anatomical structure, our method handles a finer scale by using patches. Therefore, instead of performing the fusion of nonlinearly deformed template structures, the proposed method achieves the labeling of each voxel individually by comparing its surrounding patch with patches in training subjects in which the labels of the central voxels are known. When the patch under study resembles a patch in the training subjects, their central voxels are considered to belong to the same structure, and this training patch is used to estimate the final label. By this method, several samples from each training subject can be used during the label fusion, enabling a drastic increase in the number of sample patches involved in the label estimation.

Second, multi-template methods usually use a majority voting scheme to fuse the labels (Rohlfing et al. 2004, Heckemann et al. 2006, Aljabar et al. 2009, Collins and Pruessner 2010) that considers the relevance (or weight) of all the samples labeled as similar. In the proposed method, the intensity-based distances between the patch under study and the patches in the training subjects are used to perform a weighted label fusion based on the nonlocal means estimator (Buades et al. 2005). In such an approach, the intensity-based distance between patches decreases as the relevance of the considered sample increases.

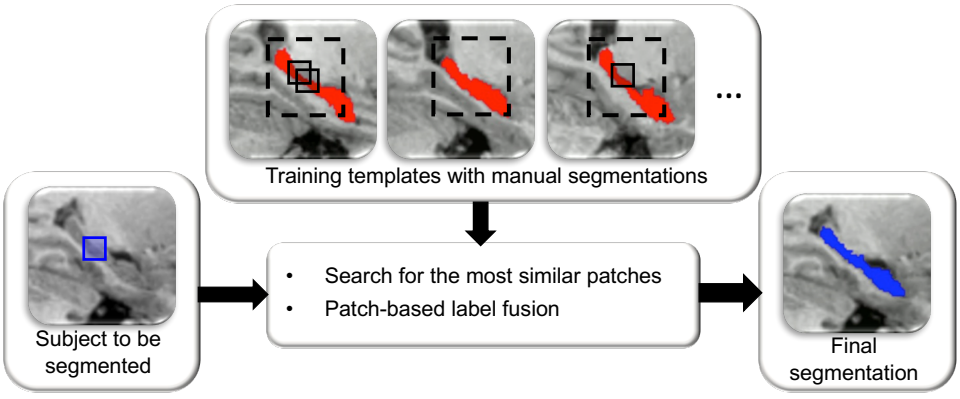

Figure 4: Overview of the patch-based segmentation method. This application of PBS to hippocampus extraction shows the main 2 steps of the method: the search of similar patches and the label fusion.

In other words, by taking advantage of the redundancy of information present in the image, the patch-based nonlocal means scheme enables the robust use of a large number of samples during estimation. This number will be significantly more important than the number of training subjects, in contrast to template-based methods (i.e., where the number of samples is the number warped subjects). Moreover, contrary to classical majority voting schemes that give the same weight to all the samples, the nonlocal means scheme enables the robust distinction of the most similar samples according to their local context (i.e., their surrounding patches). Finally, compared to atlas-based or template-based methods using nonlinear registration, the nonlocal patch-based approach has the advantage of better handling the inter-subject variability problem. Contrary to the one-to-one correspondence assumed by nonlinear warping methods between the source and the target image, the nonlocal means estimator makes it possible to deal with one-to-many mappings, which better captures the link between subjects' anatomies (see Figure 4).

1.2.3 Patch-based Segmentation Framework

To perform the patch-based label fusion step (see Eq. 3), we need first to find similar patch in the training library. In the proposed PBS method, the search of good patches within the library is designed to find the most similar patches, but is also constrained in order to avoid useless computations. Therefore, the search process uses different strategies. First, we constrain the segmentation with an initialization mask. Second, we consider the possibility that similar patches should be found in similar subjects. Then, we consider that the anatomical inter-subject variability in stereotaxic space is limited; thus, we can define a limited search volume around the location under study. This is done after preprocessing pipeline where all the subjects are linearly registered into the MNI space. Finally, we consider that two similar patches should have similar luminance and contrast. All these steps will be detailed in the following subsections.

1.2.3.1 Initialization Mask

Instead of segmenting the entire image under study, we define an initialization mask around the structure of interest. Several strategies could be used to propose an accurate initialization, such as matching the best subject (Barnes et al. 2008) followed by a morphological dilation of the mask. In our case, we chose a very fast and simple approach that uses the union of all the expert segmentations in the training database as the initial mask.

1.2.3.2 Selection of Training Subjects

A selection is performed at the subject level by selecting the N most similar training templates (Aljabar et al. 2009). In our PBS method, after linear registration of all the subjects in the MNI space, we use the sum of the squared difference (SSD) across the initialization mask. This strategy was chosen because SSD is sensitive to variation in contrast and luminance; thus, we expect to find a greater number of similar patches (in the sense of the L2-norm) in subjects with smaller SSDs. Afterwards, these N closest subjects in the training library are retained during the entire segmentation process (see Figure 5 where the three closest subjects are displayed).

1.2.3.3 Search Volume Definition

Initially, the nonlocal means denoising filter was proposed as a weighted average of all the pixels in the image (Buades et al. 2005). For computational reasons, the entire image cannot be used and the number of pixels involved has to be reduced. As done for denoising (Buades et al. 2005, Coupe et al. 2008b), we use for segmentation a limited search volume V_i defined as a cube centered on the voxel x_i under study. Thus, within each of the N selected subjects, we search for similar patches in a cubic region around the location under study (see Figure 5). This search volume can be viewed as the inter-subject variability of the structure of interest in the stereotaxic space. This variability can increase for a subject with pathology or according to the structure under consideration.

1.2.3.4 Patch Preselection

Finally, as we proposed for denoising purposes (Coupe et al. 2008b), we perform a preselection of the patches to be compared in order to reduce the computational time. By using simple statistics such as mean or variance, it is possible to discard *a priori* the most dissimilar patches. In the proposed approach, we use luminance and contrast criteria to achieve the patch preselection. Based on the first and second terms of the well-known structural similarity measure (SSIM) (Wang et al. 2004), the preselection procedure can be written as follows:

$$ss = \frac{2\mu_i\mu_{s,j}}{\mu_i^2 + \mu_{s,j}^2} \times \frac{2\sigma_i\sigma_{s,j}}{\sigma_i^2 + \sigma_{s,j}^2} \qquad \text{(Eq. 4)}$$

where μ represents the means and σ represents the standard deviations of the patches centered on voxel x_i (voxel under consideration) and voxel $x_{s,j}$ at location j in subject s. If the value of ss is greater than a given threshold, the intensity distance between patches i and j is computed. This threshold value is chosen empirically to provide a good balance between segmentation accuracy and computational time reduction.

To sum up, the proposed search enables only candidates within the most similar training subjects to be considered (SSD-based subject selection), namely, those whose locations are not too far apart in stereotaxic space (search volume) and whose local neighborhoods are similar to the neighborhood of the voxel under study (patch preselection). Hence, the introduction of outliers is limited during the nonlocal patch-based label fusion and the computational burden is drastically reduced.

1.2.3.5 Local Adaptation of Smoothing Parameter

In estimation problems using a weight function, the tuning of the decay parameter h plays a crucial role. When h is very low, only a few samples are taken into account. When h is very high, all samples tend to have the same weight and the estimation is similar to a classical average. The value of h should depend on the distance between the patch under consideration and the library content. In fact, when the library contains patches very similar to the patch under study, h needs to be decreased to drastically reduce the influence of the other patches. However, when no similar patches exist in the library, h has to be increased to relax the selection.

To achieve this local adaptation of h automatically, we propose an estimation of $h(x_i)$ based on the minimal distance between $P(x_i)$ and the considered patches $P(x_{s,j})$:

$$h^2(x_i) = \min_{x_{s,j}} \left\| P(x_i) - P\left(x_{s,j}\right) \right\|_2^2 + \varepsilon \qquad \text{(Eq. 5)}$$

This simple automatic tuning of h parameter demonstrated good performance and robustness over a large number of applications (Coupe et al. 2011, Eskildsen et al. 2012, Manjon et al. 2014, Romero et al. 2015, Giraud et al. 2016, Romero et al. 2017a). Therefore, since the search volume radius have low influence on segmentation and is fixed for computational reason, the patch radius is the only main parameter that the user has to provide to PBS method. Besides its simplicity, the very low number of hyper-parameters is one important advantage of our PBS method.

1.2.3.6 Method Overview

Figure 5 presents an overview of the different steps used to segment one voxel x_i included in the initialization mask. After the selection of the N most similar subjects in the training library ($N = 3$ in this example), the patch $P(x_i)$ (in green) is compared with all the patches $P(x_{s,j})$ contained in the search volume V_i within the N selected subjects. The most similar patches $P(x_{s,j})$ (in blue) to the patch $P(x_i)$ obtain the highest weights, as shown in the weight maps. For the 2D slice in this illustration, 12 labeled samples have significant weights in subject s_1, the two most similar patches are in subject s_2, and no similar patches are found in subject s_3.

In (Coupe et al. 2010a, Coupe et al. 2011), we showed that accurate segmentations of anatomical structures can be obtained using this simple patch-based label fusion framework. Compared to multi-template segmentation methods, the main advantages of our PBS are to be faster and more robust since we do not perform any nonlinear registration step that is a difficult task, subjects to failures, and that is computationally demanding. The high segmentation quality, the good robustness and the simple tuning of our PBS participated to its success over the last years. It is now well-established that PBS obtained similar or better performance than template-warping methods with reduced computational time and lower pipeline failures (Zandifar et al. 2017). However, despite its qualities, our original PBS framework had some limitations that we tried to address by proposing improvements and extensions.

First, although the computational time is drastically reduced compared to frameworks based on multiple nonlinear registrations, the original PBS method was too slow for large structure segmentation such as brain or for processing large database. Therefore, we proposed multiresolution propagation (Eskildsen et al. 2012, Manjon et al. 2014) and ultrafast patch search (Ta et al. 2014, Giraud et al. 2016).

Second, for structures with low contrast boundaries, the high degree of freedom of PBS may lead to suboptimal results. To overcome this limitation, we proposed several regularization strategies to improve segmentation quality using level-set model (Hu et al. 2014), patchwise aggregation (Manjon et al. 2014) or by combining PBS and multiple nonlinear registrations strategies (Hu et al. 2014, Romero et al. 2017a, Zandifar et al. 2017).

Finally, brain structures boundaries are not always visible on the same MRI modality. Therefore, for some specific structures such as white matter lesions (Guizard et al. 2015a), hippocampus subfields (Romero et al. 2016) or sub-thalamic nucleus (Haegelen et al. 2013), we proposed to take advantage of the complementary of MRI modalities.

Some of these improvements will be presented in the next section. We will show that our original PBS can be optimized to propose near real time segmentation framework and to significantly improve its segmentation quality.

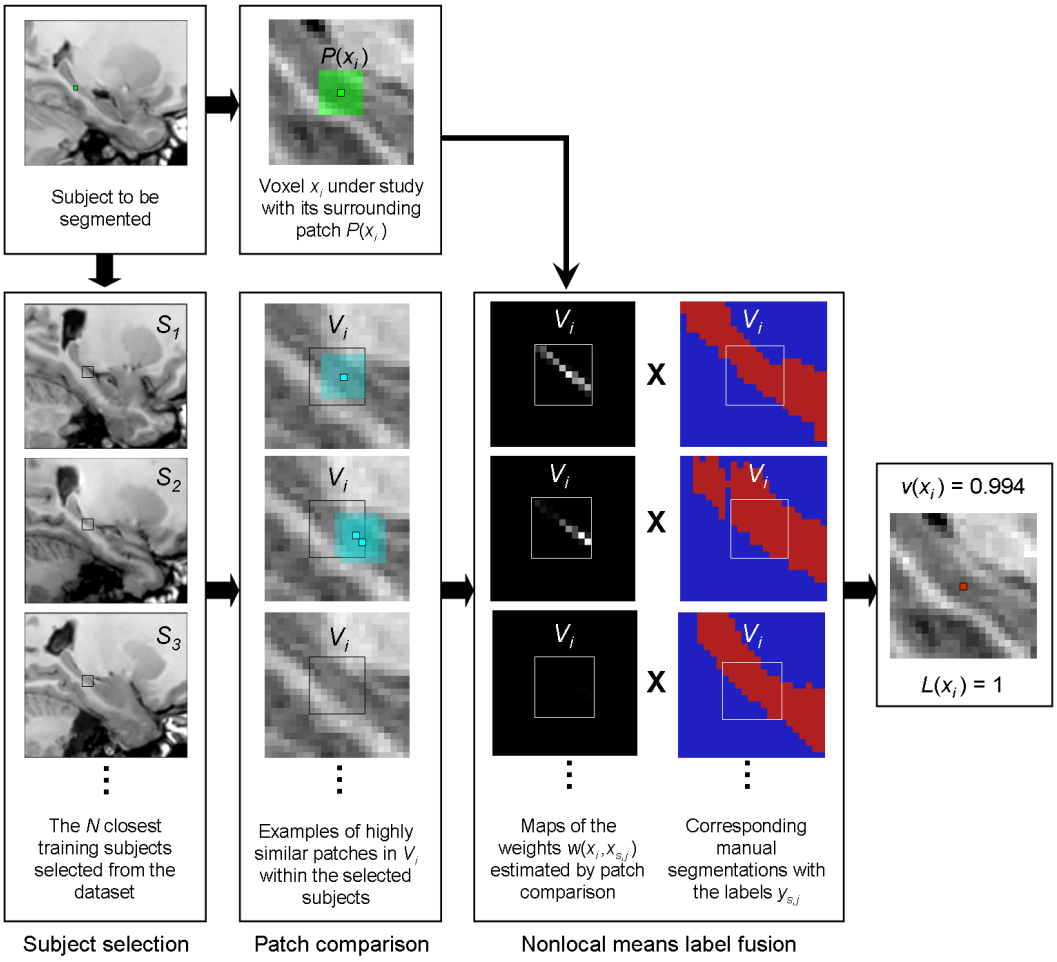

Figure 5: Overview of the different steps involved in achieving PBS of one voxel x_i included in the initialization mask. The patch $P(x_i)$ (in green) is compared with all the patches $P(x_{s,j})$ contained in the search volume V_i within the N selected subjects ($N = 3$ in this example). The weight maps show that the highest weights are obtained by the most similar patches $P(x_{s,j})$ (in blue) to the patch $P(x_i)$. After the nonlocal means fusion of the expert-based labels $y_{s,j}$, the resulting estimation is $v(x_i) = 0.994$. Thus, the final label is $L(x_i) = 1$.

1.3 Extensions

Since its introduction in (Coupe et al. 2010a, Coupe et al. 2011), PBS has been the focus of many attentions. This method is now considered as a *state-of-the-art* method in the field of anatomical structure segmentation. Over the last years, we proposed several strategies to improve its segmentation quality such as using dictionary learning and sparse coding (Tong et al. 2013), patchwise aggregation (Romero et al. 2015), multiscale strategy (Eskildsen et al. 2012, Manjon et al. 2014), multi-feature comparison (Giraud et al. 2016) or multimodal extension (Guizard et al. 2015a, Romero et al. 2016, Romero et al. 2017b). Moreover, we showed that higher segmentation accuracy can be obtained by combining PBS with active appearance model (Hu et al. 2014) and/or nonlinear registration (Hu et al. 2014, Romero et al. 2017a). Finally, we proposed a very fast implementation by adapting the PatchMatch algorithm (Barnes et al. 2009) to PBS problem (Ta et al. 2014, Giraud et al. 2016). In this section, we will present only few of them that we consider as major contributions.

1.3.1 Multiresolution Propagation Framework

As previously mentioned, the original PBS is not well-suited for large structure such as brain in terms of computational time and segmentation regularity. Indeed, PBS method cannot be directly applied to brain extraction, because false positives are likely to occur as extra-cerebral tissue since non brain tissue may resemble to brain at the patch scale. Moreover, the computational complexity is high and this becomes a significant problem for large structures. Therefore, in (Eskildsen et al. 2012), we presented the adaptation of our PBS approach to perform brain extraction by proposing to apply the PBS within a multiresolution propagation approach.

To obtain optimal performance for brain extraction, the patch size needs to be large compared to the patch sizes used to segment smaller structures such as the hippocampus. For example, a small patch in the dura may look like gray matter in a T1-weighted (T1w) MRI. Thus, a large patch size, including more structural information, is needed to avoid inclusion of extra-cerebral tissue, such as dura or fat. This is computationally impractical at the stereotaxic resolution. Therefore, we suggest integrating the patch-based segmentation within a multiresolution propagation framework, which provides the opportunity to effectively have spatially large patch sizes while still being computationally practical.

The multiresolution propagation framework enables propagation of segmentation across scales thanks to a pyramidal approach by using the resulting segmentation at the previous scale to initialize the segmentation at the current scale. The library images, labels, initialization mask, and target image at the stereotaxic resolution are all resampled to a lower resolution, before performing the patch-based segmentation. The nonlocal means estimator at the previous resolution is propagated to a higher resolution using interpolation (see Figure 6). The nonlocal means estimator function $v(x_i)$ (see Eq. 3) can be considered as the confidence level of which label to assign the voxel. Values close to 0 are likely background, while values close to 1 are likely object. We define a confidence level α to assign labels to the voxels at each scale. Voxels with $v(x_i)<\alpha$ are labelled background, and voxels with $v(x_i)>(1-\alpha)$ are labelled structure. Segmentation of these two sets of voxels is considered final, and they are excluded from further processing. Voxels with $v(xi)$ in the range $[\alpha; 1-\alpha]$ are propagated and processed at a higher resolution.

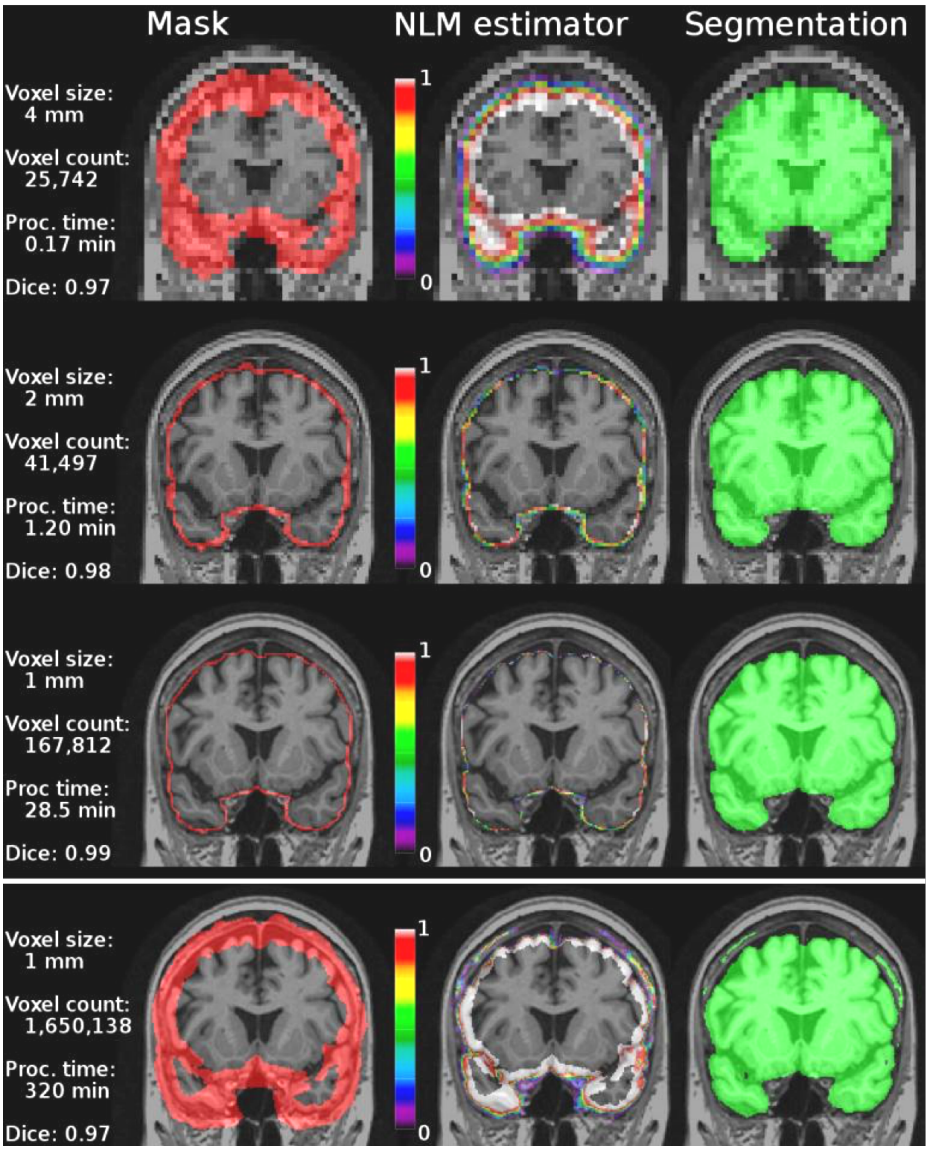

Figure 6: The multiresolution propagation segmentation process (row 1-3) compared to a single resolution approach (row 4). Column 1: Initialization mask. Column 2: Nonlocal means (NLM) estimator map. Column 3: Segmentation by thresholding the NLM estimator and adding the intersection mask. Processing times are accumulated time from initialization. Notice the inclusion of dura in the single resolution approach.

This procedure is repeated until the resolution of the stereotaxic space is reached. In this manner, the initialization mask of each resolution step is limited to the voxels with uncertain segmentation at the previous step. At the stereotaxic resolution, final segmentation is done by thresholding the estimator $v(x_i)$ at 0.5 as in the original PBS. The proposed multiresolution framework greatly reduces the computational cost as assessed in Figure 6. In addition, it is interesting to note that the proposed multiscale method enables to improve the segmentation accuracy and to limit introduction of outliers. This aspect will be discussed in the results section.

1.3.2 Fast Search based on PatchMatch Algorithm

Beside multiresolution strategy, PBS method can be accelerated by improving the search of similar patches. In (Ta et al. 2014, Giraud et al. 2016), we introduced a new Optimized PAtchMatch for Label fusion (OPAL) to address this issue. Originally, the PatchMatch (PM) (Barnes et al. 2009) algorithm was introduced to efficiently find patch correspondences between two 2D images. For each patch within the first image, an approximate nearest neighbor (ANN) is found within the second image. The algorithm is based on a cooperative and randomized strategy resulting in very low computation time and enabling near real-time processing. In (Giraud et al. 2016), we investigated the use of the PM for patch-based anatomical structures segmentation using multi-template training library. Thanks to our Optimized PM algorithm, OPAL can produce segmentations in a drastically reduced computation time compared to original PBS method.

1.3.2.1 PatchMatch Algorithm

The original PM algorithm (Barnes et al. 2009) is a fast and efficient approach that computes patch correspondences (matches) between two 2D images (e.g. S and T). The key point of this method is that good matches can be propagated to the adjacent patches within an image. This propagation, combined with random matches, leads to a very fast convergence with limited computational burden. The core of the algorithm is based on three steps: initialization, propagation, and random search (see Figure 7). The initialization consists in randomly associating each patch of S with a corresponding patch in T, to obtain an initial ANN field. The two following steps are then performed iteratively to improve the ANN field. The propagation step uses the assumption that when a patch $P(x_i) \in S$ matches well with a patch $P(x_j) \in T$, then the adjacent patches of $P(x_i) \in S$ should match well with the adjacent patches of $P(x_j) \in T$. Next, the random search step consists of a random sampling around the current ANN to escape from local minima.

1.3.2.2 Optimized PatchMatch for Label Fusion

In contrast to (Barnes et al. 2009) where two 2D images are considered, OPAL finds the patch correspondences between the subject to be segmented and a training library of 3D templates. One advantage of the PM algorithm is that its complexity only depends on the size of image S to process and not on the size of the compared image T, (i.e., the size of the training library in the OPAL case). This important fact enables OPAL to consider the entire training library without any template preselection step at constant complexity in time. Moreover, for each patch in the subject to be segmented, OPAL computes the best k-ANN matches and not only one match as done in (Barnes et al. 2009). The OPAL algorithm is explained in detail in the Figure 7 that proposes a schematic overview. To clearly illustrate our Optimized PatchMatch (OPM) key steps, only three training templates are considered, two iterations are performed and 3D MRI volumes are displayed in 2D.

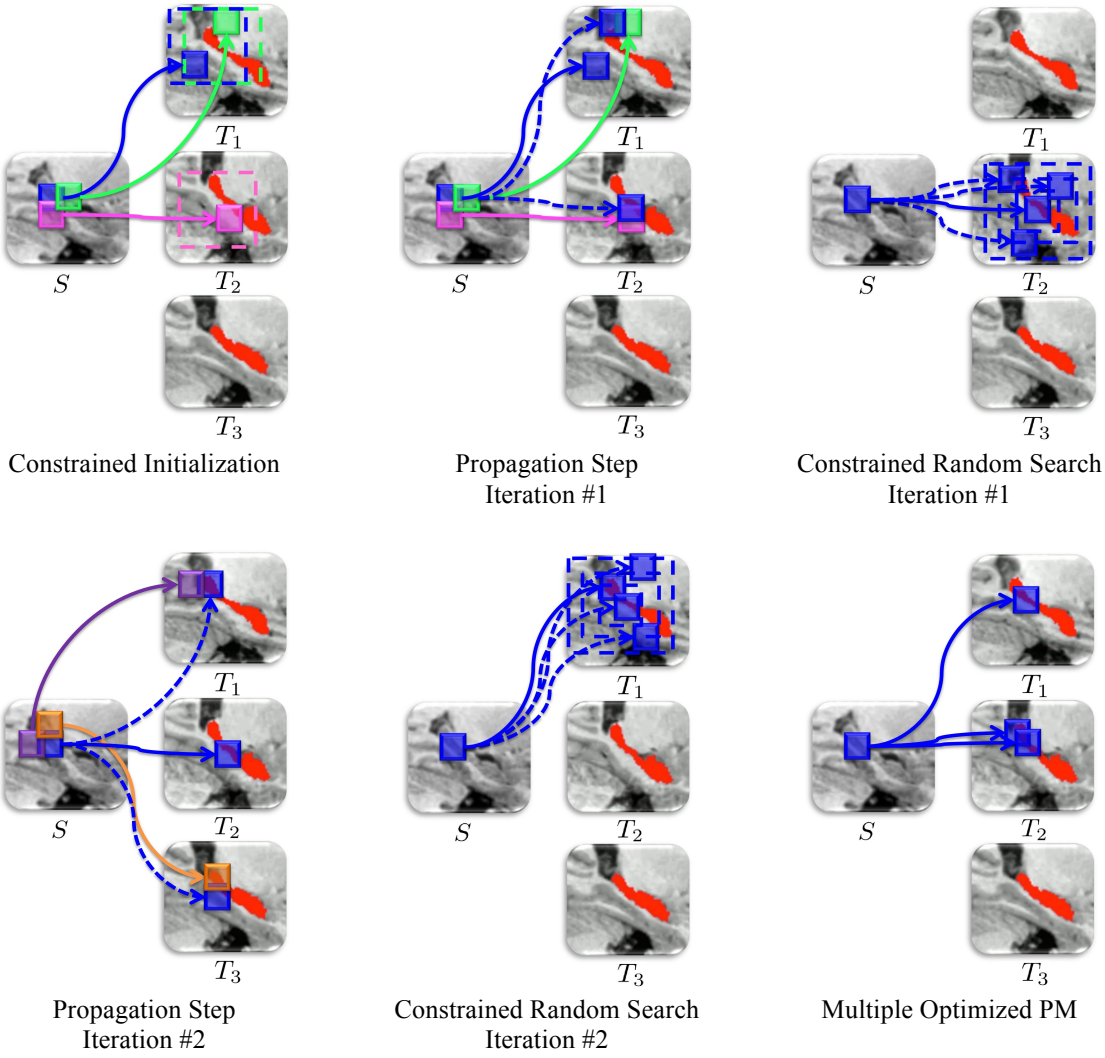

Figure 7: Optimized PatchMatch (OPM) main steps. In this figure, the representation of OPM steps focuses on the blue patch in S. Green, pink, purple and orange colors represent the adjacent patches of the blue patch. During the constrained initialization, patches of the subject S are matched (full lines) to a random patch of the library within an initialization search window (three are displayed). The propagation step, is represented for iteration #1 and #2. The shifted correspondences of recently processed adjacent patches are tested for improvement (dotted lines). Constrained random search for iteration #1 and #2 are represented for the blue patch. Random tests are performed within a decaying search window around the current best match, within the current best template. Finally, the result of multiple independent ANN searches by OPM is illustrated.

1.3.2.3 Constrained Initialization

In the original paper of PM (Barnes et al. 2009), the initialization consists in assigning, for each patch of *S*, a random correspondence which can be located everywhere in *T*. In the case of multi-template method based on 3D MRI, the natural extension of this initialization step is to assign, for each patch of the 3D image of the subject to segment a random patch correspondence in the template library. However, as we deal with linearly registered MRI volumes, we proposed to constrain the random initial position to be within a fixed search window centered around the current voxel position. Then, for each voxel in the subject to be segmented, a template index is assigned using i.i.d. random variable. Consequently, each patch in the subject to be segmented is associated to a unique random match among all templates of the training library (see Figure 7). This constraint has two advantages. First, it improves the matching convergence, making good use of the linear registration between training templates and the subject. Second, limiting the initialization to a fixed window prevents the algorithm from finding similar

patches in terms of intensity (low SSD) that are spatially far, leading to potential segmentation errors. As a consequence, our constrained initialization reinforces spatial proximity between voxels in the subject to be segmented and their matches in the training library and makes the algorithm converge faster. As in the original PM algorithm, after this constrained initialization, propagation and random search steps are performed iteratively to improve the patch correspondence.

1.3.2.4 Propagation Step

The propagation step of our optimized PM is the simple 3D extension of the one proposed in (Barnes et al. 2009). For each patch, an ANN improvement is performed by testing if the shifted ANN of its 6 directly adjacent patches provides a better match. Figure 7 illustrates this step, where the blue dotted lines correspond to the test of shifted adjacent neighbors in training library, to improve the current blue patch correspondence. In the example showed in Figure 7, the best match for the blue patch moves from the template T_1 to T_2 at iteration #1 and from T_2 to T_1 at iteration #2. The propagation step is a core stage of the OPAL algorithm since it allows a patch correspondence to move over all the templates in the training library. Thus, the ANN of the current voxel can move from one template to another one, since the ANN of the adjacent voxels are not necessarily in the same template.

1.3.2.5 Constrained Random Search

In the original PM algorithm, the random search step is performed on all dimensions. In contrast to the original method, OPAL deals with a library of images. Therefore, we modify the random search step to take into account this aspect. In order to ensure spatial consistency, OPAL performs the random search only in the current template containing the current best patch correspondence within a decaying search window. The process stops when the window is reduced to a single voxel. The decaying search window size is empirically defined as the size of the initialization window. Figure 7 presents examples of such fixed template random search where the decaying search windows are represented in dotted blue lines.

1.3.2.6 Multiple PM and Parallel Computation

Finally, contrary to (Barnes et al. 2009) that only estimates the best match with PM, OPAL computes k-ANN matches. These ANNs are then used to perform the PBS as described in previous section. In the literature, an extension of the original PM algorithm to k-ANN case has been proposed in (Barnes et al. 2010). The suggested strategy is to build a stack of the best visited matches. At each new tested match, the distance is compared to the one of the worst ANN among the stack. If there is an improvement in terms of SSD, the worst ANN is replaced by the new match. However, to parallelize such an approach, the current image must be split into several parts. Since PM uses propagation of good matches between adjacent patches, any split would lead to boundary issues. Therefore, in OPAL, we decided to implement the k-ANN search through k independent OPM, denoted as k-OPM. This leads to a more efficient and simple multi-threading. Consequently, each thread can run an OPM without any dependencies to the other ones. Figure 7 illustrates the result of the multiple OPM steps with k = 3. One can note that k independent OPM can lead to the same ANN for a given voxel. The redundancy of the same ANN in the ANN map is not an issue, since each contribution is weighted during the patch-based label fusion step (see Eq. 3).

With OPAL, PBS method can achieve structure segmentation (e.g., hippocampus) in less than 2s with a standard computer as we will present it in the results section.

1.3.3 Multiscale and Multifeature Framework

Due to the high computational cost of previously published multi-template methods, most were designed in a monoscale and monofeature context. Thanks to the computational efficiency of OPAL, we proposed a new framework to simultaneously perform multiscale and multifeature analysis with late aggregation of estimators (Giraud et al. 2016). Figure 8 illustrates the proposed multifeature and multiscale framework.

1.3.3.1 Multiscale Estimators

As previously mentioned with our multiresolution propagation framework for brain extraction, the structure analysis highly depends on the size of the patch in PBS method. The patch size needs to be large enough to capture the local geometry and to prevent discontinuities in the segmentation. However, using very large neighborhoods may reduce the probability of finding similar patches in the library. Although the optimal patch size can be determined by experiments for a given dataset, multiscale approaches may significantly improve segmentation accuracy. In (Eskildsen et al. 2012), the idea was to perform multiresolution propagation to reduce computational time and to limit false positive during the segmentation of large structures. In (Giraud et al. 2016), by taking advantage of the OPAL speed, we proposed a different multiscale strategy where fully independent multiscale searches are performed with patches of different size at the highest resolution. This multiscale framework is not necessary designed for large structures but for improving segmentation accuracy and to limit the impact of patch size choice on segmentation accuracy. While in (Eskildsen et al. 2012) we proposed a multiresolution propagation of label confidence, in (Giraud et al. 2016) we proposed a fusion of estimator maps where multiscale refers to the simultaneous use of patches of different sizes, and where the images are considered with their initial resolution (see Figure 8).

1.3.3.2 Multifeature Estimators

Similarly, the search for similar patches by OPAL can also be carried out independently on different features (edges, textures, etc.). In (Giraud et al. 2016), we showed that using the gradient norm in addition to the original MRI intensities increases the segmentation accuracy. Therefore, we proposed to apply OPAL at different scales for each considered feature (see Figure 8). The resulting estimator maps are then merged *a posteriori* using late fusion approach (i.e., a simple average in our case). The fusion of all the maps obtained through the independent searches improves the diversity of the selected patches and produce better segmentation quality, as shown in the results section.

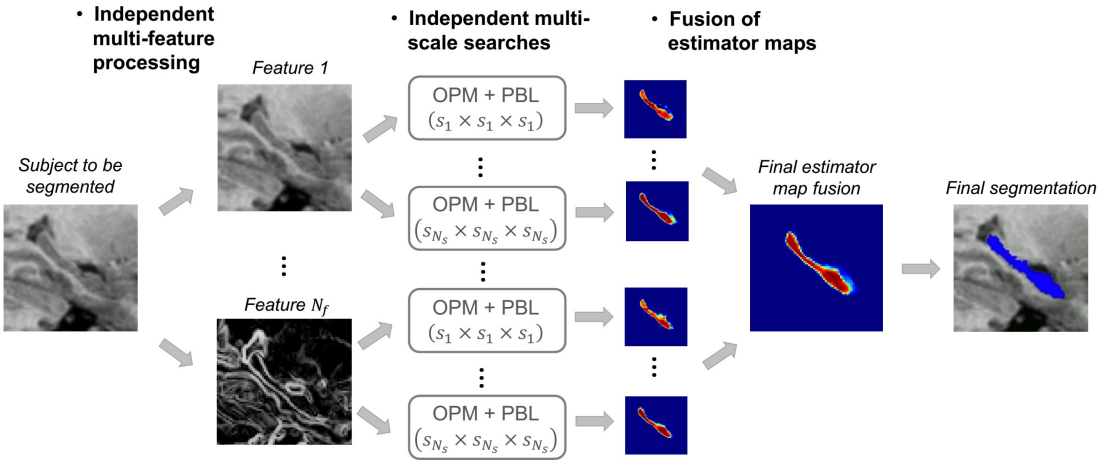

Figure 8: Fusion of multifeature and multiscale label estimator maps. Our method is applied with different patch sizes, on different features independently. At the end, all the estimator maps are fused to provide the final segmentation.

1.3.4 Multimodal Extension

Over the past years, we also investigated the use of multimodal framework to take advantage of information derived from different MRI modalities (Haegelen et al. 2013, Guizard et al. 2015a, Romero et al. 2016, Romero et al. 2017b). Contrary to the original patch-based segmentation method (Coupe et al. 2011) using a single contrast library (e.g., only T1w MRI), we proposed an adaptation of the PBS method to take advantage of multi-contrast images. In Multiple Sclerosis (MS) for instance, multi-contrast manual and automatic segmentation methods have shown to improve the identification of MS lesions. In addition, for hippocampus subfields segmentation, the use of multi-contrast information is also crucial to improve subfields borders detection. Therefore, in (Guizard et al. 2015a) and (Romero et al. 2016, Romero et al. 2017b), we adapted the nonlocal means weighting function (see Eq. 1) for various contrasts (m) as follows:

$$w_M(x_i, x_{s,j}) = e^{-\left(\sum_{m \in M}\left(\frac{\left\|P_m(x_i) - P_m(x_{s,j})\right\|_2^2}{h_m^2}\right)\right)} \quad \text{(Eq. 6)}$$

Here m represents the different MRI modalities. The smoothing parameter h_m is estimated for each considered contrast independently (i.e., the per contrast minimum distance as defined in Eq. 5). Moreover, the L2-norm distance is estimated between patches of the same modality. This simple extension of the weight estimation enables to take advantage of the additional information derived from different MRI sequences. Therefore, w_M is the multimodal weight estimated over the M available modalities. For instance, the commonly modalities used in MS lesion segmentation are T1w, T2w or FLAIR while for hippocampus subfields segmentation T1w and T2w are usually used. Finally, the use of the minimal distance for the automatic tuning of h_m enables to automatically homogenize the importance of each modality intendedly of their intensity range or contrast.

1.3.5 Combination of PBS, Nonlinear Registration and Error Correction

Finally, we also investigated the combination of PBS and multiple nonlinear registrations framework (Fonov et al. 2012, Hu et al. 2014, Romero et al. 2017a, Zandifar et al. 2017). The main idea is to take benefit from the high flexibility of PBS to capture local geometry and to take advantage from implicit shape priors provided by multi-template warping methods (Collins and Pruessner 2010). To integrate global shape constraints into the segmentation and increase the local structure fitting, we developed several fully automatic segmentation methods that combined multi-template warping method and patch-based technique into a general two-stages segmentation framework. In the first-stage segmentation, the multi-template warping is used to capture the statistical characteristics of shape and intensity information in the training data. Although the multi-template warping method take into account local geometry, its ability to recover fine details at structure borders is limited by the regularization of the nonlinear registration. Thus, there is often some "blurring" of the structure shape. This issue can be addressed by using PBS technique, which is employed as a second-stage segmentation to locally refine the segmentation resulting from the first-stage segmentation. In (Hu et al. 2014), we proposed to perform a segmentation based on multi-template warping and then to refine this segmentation thanks to the original PBS. In (Romero et al. 2017a), we proposed first to nonlinearly register all the training templates to the subject under study and then to apply OPAL over this subject-specific training library. Finally, in (Zandifar et al. 2017), we proposed to combine PBS (Coupe et al. 2011), multi-template warping (Collins and Pruessner 2010) and systematic error correction based on machine learning (Wang et al. 2011). In (Wang et al. 2011), the authors presented a method that learns the pattern of mismatch between automatic segmentation labels and their corresponding manual segmentations. This correction method uses a classification technique in which the classifier is trained by a set of automatically segmentation and their corresponding manual segmentation. For new test image, the method corrects the mislabeled segmentation using the learned pattern. Both intensity and neighborhood information are used as features to train an AdaBoost learner (Freund and Schapire 1995). In (Romero et al. 2017b), we proposed a more efficient error correction based on neural network. At present, our last pipelines are all based on the combination of OPAL (Giraud et al. 2016), multi-template warping (Heckemann et al. 2006, Collins and Pruessner 2010) and systematic error correction (Wang et al. 2011). According to our experience, this 3-stage approach enables fast and accurate segmentation.

1.4 Results

In this section, we will present main results obtained on different structures and for several clinical applications. First, we will compare PBS with the proposed improvements for hippocampus (HIPP) segmentation, a key structure involved in many neurological pathologies (Coupe et al. 2011, Tong et al. 2013, Giraud et al. 2016). Second, we will propose a comparison of our brain extraction based on multiresolution propagation PBS with *state-of-the-art* methods (Eskildsen et al. 2012, Manjon et al. 2014). Then, we will present results obtained with the combination of PBS and nonlinear registration for cerebellum lobules and HIPP subfields segmentation (Romero et al. 2016, Romero et al. 2017a, Romero et al. 2017b). Finally, we will show the interest of the proposed multimodal framework in the context of MS lesions segmentation (Guizard et al. 2015a).

1.4.1 Hippocampus Segmentation

In (Giraud et al. 2016), we compared several methods to segment the hippocampi of healthy subjects. The HIPP plays an important role in human memory and orientation. Moreover, HIPP dysfunction is involved in a variety of diseases, including Alzheimer's Disease (Jack et al. 2000), post-traumatic stress disorder (Bremner et al. 1995), major depression (Bremner et al. 2000), schizophrenia (Tanskanen et al. 2005, Buss et al. 2007), and epilepsy (Bernasconi et al. 2003). This structure is especially difficult to segment because of its small size, high variability, low contrast, and discontinuous boundaries in MR images (Chupin et al. 2007, Siadat et al. 2007). Finally, the HIPP is composed of several tissue types, which prevents the use of simple intensity-based techniques.

The HIPP dataset used for validation consists of T1w MR images of 80 subjects randomly extracted from a group of 152 young, healthy individuals acquired on a 1.5T MRI scanner in the context of the International Consortium for Brain Mapping (ICBM) project (Mazziotta et al. 1995). The MR images were manually segmented by an expert directly into the stereotaxic space. For each subject, the HIPP label was manually defined using the protocol described by Pruessner et al. (2000). The resulting segmentations obtained an intraclass reliability coefficient (ICC) of 0.900 for inter-rater reliability (4 raters) and 0.925 for intra-rater reliability (5 repeats). In the following, for each dataset, a leave-one-out procedure was performed. The kappa index (Dice coefficient or similarity index) (Zijdenbos et al. 1994) was then computed by comparing the expert segmentations with those obtained with our methods. For two binary segmentations A and B, the kappa index was computed as:

$$\kappa(A,B) = \frac{2|A \cap B|}{|A| + |B|}. \qquad \text{(Eq. 7)}$$

As usual in quantitative MR analysis, manual segmentation is considered the gold standard (Pruessner et al. 2000) and thus used as reference.

*Table 1: Comparison of different approaches in terms of segmentation accuracy and computation time for hippocampus segmentation. The compared methods are: our original PBS (Coupe et al. 2011), a multi-template wrapping method based on multiple nonlinear registrations (Collins and Pruessner 2010), our PBS using sparse coding and dictionary learning to improve weight estimation (Tong et al. 2013) and our PBS based on optimized PatchMatch called OPAL (Giraud et al. 2016). * indicate significantly better results at p<0.01 compared to dictionary learning.*

Methods	DICE	Computational time
PBS (Coupe et al. 2011)	88.2	662s (x700)
Multi-template warping (Collins and Pruessner 2010)	88.6	3974s (x4300)
Sparse coding (Tong et al. 2013)	88.7	5587s (x6000)
Dictionary learning (Tong et al. 2013)	89.0	943s (x1000)
OPAL (Giraud et al. 2016)	**89.9***	**0.92s**

Table 1 presents a comparison of our original PBS with different improvements that we proposed over the past years and a multi-template warping used as reference. Compared to the original PBS (Coupe et al. 2011), OPAL improves segmentation accuracy by 1.7 percentage points (pp) thanks to multiscale and multifeature framework while being 700× faster thanks to the optimized PatchMatch search. Compared to the most accurate method on this dataset based on our patch-based dictionary learning (Tong et al. 2013), OPAL obtained higher Dice coefficients with a p-value inferior to 10^{-12} (obtained from a paired t-test on the OPAL and Dictionary learning sets of Dice coefficients) for a computation times 1000× faster. These results show the significant improvement of segmentation quality obtained when using multiscale and multifeature PBS.

1.4.2 Brain Extraction: BEaST

In (Eskildsen et al. 2012), we compared our multiresolution propagation framework for brain extraction (called BEaST) with two *state-of-the-art* methods – BET based on deformable model (Smith 2002) and VBM8 based on tissue classification (http://dbm.neuro.uni-jena.de/vbm/download). To validate these methods, we performed a Leave-One-Out Cross Validation (LOOCV) using three datasets: the NIH-funded MRI study of normal brain development (termed here the NIH Pediatric Database, or NIHPD) (Evans and Group 2006) (age: 5–18y), the International Consortium for Brain Mapping (ICBM) database (Mazziotta et al. 1995) (age: 18–43y), and the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Mueller et al. 2005) (age: 55–91y). The NIHPD and ICBM databases consisted of healthy subjects, while the ADNI database, in addition to cognitive normal (CN) subjects, contained scans of subjects with AD and mild cognitive impairment (MCI). This way, almost the entire human life span was covered and subjects with atrophic anatomy were included.

Table 2 presents the comparison of BET, VBM8 and our multiresolution PBS segmentation method. BET yielded very high DICE for ICBM and NIHPD, while the results are more mixed on ADNI. VBM8 provided slightly lower DICE on ICBM and NIHPD. On the ADNI dataset, VBM8 provided on average DICE values larger than those obtained by BET and is more consistent in its segmentation. Our multiresolution PBS yielded consistently and significantly higher DICE on all datasets.

*Table 2: Comparison of different approaches in terms of DICE for brain segmentation. The compared methods are: the Multiresolution PBS (Eskildsen et al. 2012), BET (**S**mith 2002) and VBM8. * indicate significantly better results at p<0.01.*

Methods	ICBM	NIHPD	ADNI
BEaST (Eskildsen et al. 2012)	**99.0***	**98.1***	**98.5***
BET (Smith 2002)	97.5	97.5	94.4
VBM8	96.7	97.2	96.3

Figure 9 shows typical examples of brain masks obtained by BET, VBM8 and our multiresolution PBS on the five different groups tested here (NIHPD, ICBM, ADNI-CN, ADNI-MCI, ADNI-AD). On NIHPD and ICBM data, BET behaved quite well with only minor segmentation errors, such as inclusion of the transverse sinus and part of the eye sockets. On ADNI data, more serious errors were found using BET. These include inclusion of dura and marrow of the skull while gyri are often cut off in atrophic brains. VBM8 had a tendency to perform over-segmentations on all groups and sometimes included dura proximate to the brain, carotid arteries, ocular fat / muscle, and parts of the eyes. On the positive side, VBM8 rarely removes part of the brain due to the consistent over-segmentation. BEaST generally provided a more consistent and robust segmentation without serious errors.

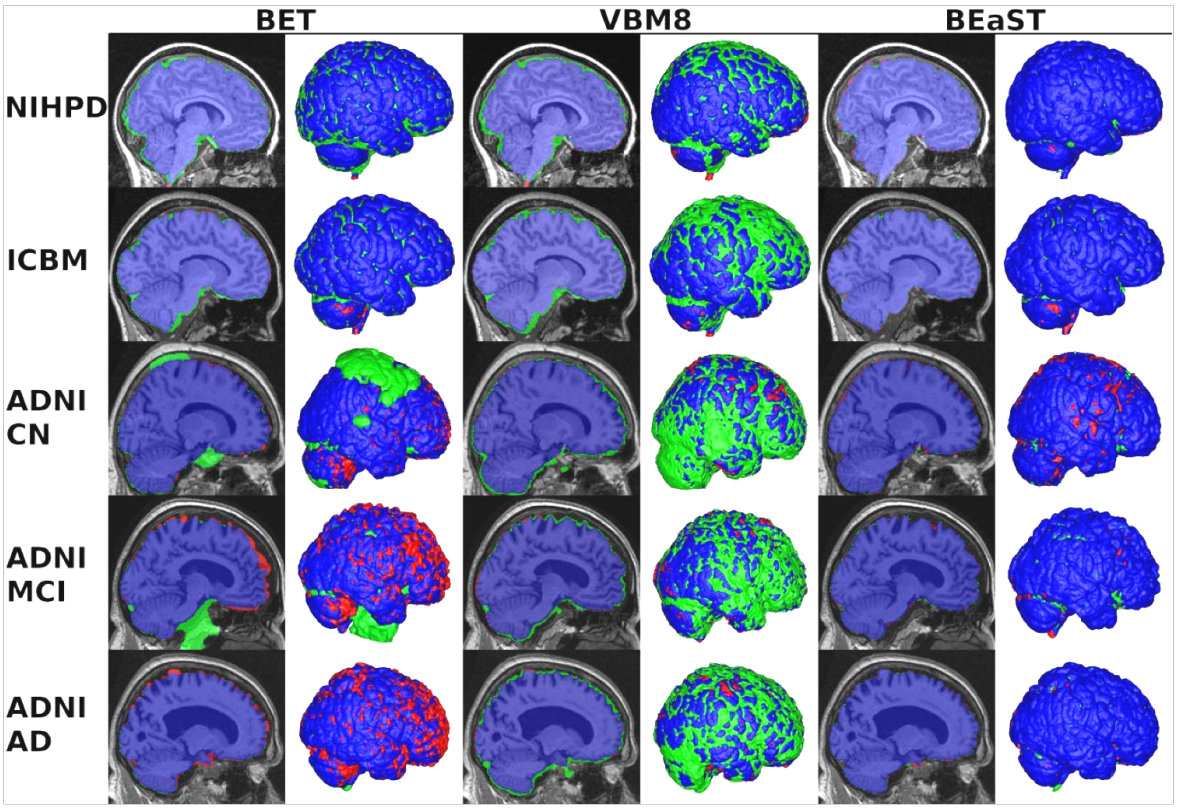

Figure 9: Typical results using BET, VBM8 and our multiresolution PBS called BEaST on the five test groups. The figure shows sagittal slices and 3D renderings of the segmentations. Blue voxels are overlapping voxels in the segmentation compared to the gold standard. Green voxels are false positives and red voxels are false negatives.

1.4.3 Lobules and Subfields Segmentation

1.4.3.1 Cerebellum Lobules Segmentation: CERES

In (Romero et al. 2017a), we have introduced a new pipeline for cerebellum lobule segmentation that was based on an adaptation of OPAL (Ta et al. 2014, Giraud et al. 2016) called CERES. The human cerebellum is a neuroanatomical structure within the human brain located below the cerebrum and connected to the brainstem through the cerebellar peduncles. Although it represents a small percentage of the total intracranial volume, about 10%, it plays a key role in motor coordination and learning. Cerebellar anatomy consists of a white matter tree structure located behind the pons. It is divided into two hemispheres (left and right), with each white matter branch surrounded by a layer of grey matter that creates folds called foliations (see Figure 10). These grey matter folds are denominated cerebellum lobules. The size, position and number of (visible) lobules is highly variable between subjects which makes the segmentation process challenging.

The proposed method CERES is based on OPAL that has been adapted to segment the cerebellum anatomy using a library of nonlinearly registered cases instead of using only an affine registration as in the original PBS method. CERES was compared with 3 *state-of-the-art* methods: SUIT a single atlas method (Diedrichsen 2006), MAGeT a multi-template method (Chakravarty et al. 2013, Park et al. 2014) and RASCAL a multi-template method combined with PBS (Weier et al. 2014). These methods were compared using high resolution MR images from 5 healthy volunteers (2 males, 3 females, aged 29–57). These high resolution MR images were manually segmented by two expert raters that allowed for both inter- and intra-rater comparisons of segmentations to validate the consistency of the manual segmentation protocol as defined in (Park et al. 2014). The cerebellum was manually segmented into 26 structures: White matter and Lobules I-IV, V, VI, Crus II, VIIB, VIIIA, VIIIA, and X considering left and right hemispheres.

*Table 3: Mean and standard deviation for DICE coefficient values for SUIT a single atlas method (Diedrichsen 2006), MAGeT a multi-template method (Chakravarty et al. 2013), RASCAL a multi-template method with PBS refinement (Weier et al. 2014) and the proposed CERES (Romero et al. 2017a). Best results are marked in bold (significant differences ($p<0.05$) are marked with * for SUIT, † for MAGeT and ‡ for RASCAL comparison).*

Structure	SUIT (Diedrichsen 2006)	MAGeT (Chakravarty et al. 2013)	RASCAL (Weier et al. 2014)	CERES (Romero et al. 2017a)
Lobule I-IV	0.7435 ± 0.0880	**0.8055** ± 0.0964	0.7703 ± 0.1107	0.7898 ± 0.1021 *
Lobule V	0.6598 ± 0.1026	0.7429 ± 0.1369	0.6730 ± 0.1560	**0.7561** ± 0.1332 *‡
Lobule VI	0.7800 ± 0.0543	**0.8762** ± 0.0365	0.7994 ± 0.0523	0.8695 ± 0.0316 *‡
Lobule Crus II	0.7430 ± 0.0631	0.7787 ± 0.0678	0.7300 ± 0.0667	**0.8096** ± 0.0569 *‡
Lobule VIIB	0.5701 ± 0.1572	0.6013 ± 0.1476	0.5761 ± 0.1137	**0.6850** ± 0.1205 *‡
Lobule VIIIA	0.7134 ± 0.0996	0.7330 ± 0.0928	0.6701 ± 0.1426	**0.7926** ± 0.0759 *†‡
Lobule VIIIB	0.7721 ± 0.0596	0.8012 ± 0.0607	0.7654 ± 0.0931	**0.8533** ± 0.0390 *†‡
Lobule X	0.6955 ± 0.0512	**0.7721** ± 0.0475	0.7275 ± 0.0680	0.7548 ± 0.0469 *
Average	0.7097 ± 0.0689	0.7639 ± 0.0792	0.7140 ± 0.0487	**0.7888** ± 0.0409 *‡

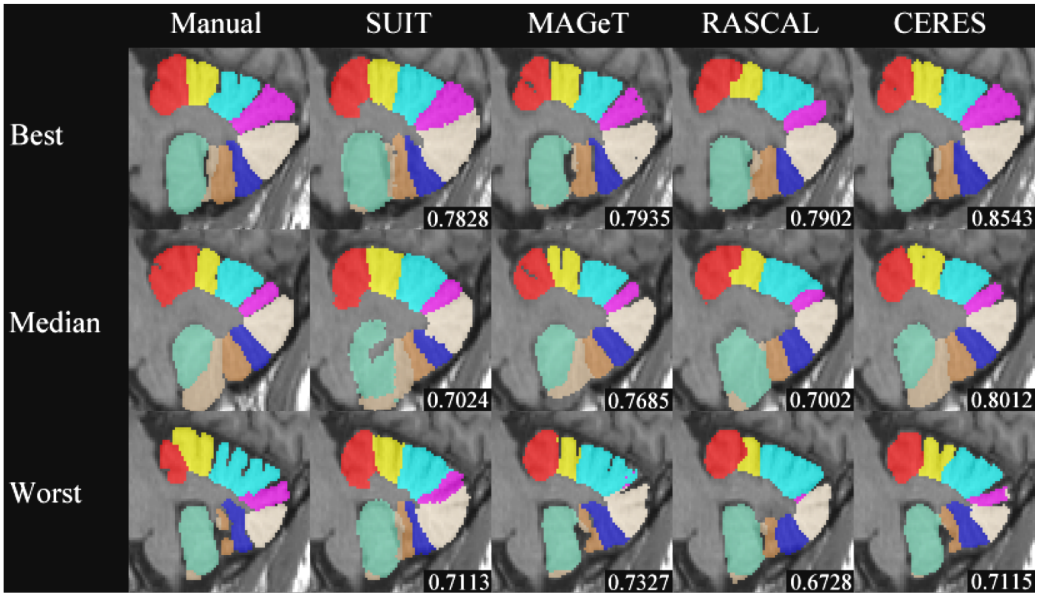

Figure 10: Examples of manual, SUIT, MAGeT, RASCAL and CERES segmentation over the cases with the best, median and worst DICE for CERES. Corresponding DICE is provided for each segmentation.

In Table 3, DICE coefficient values for the different compared methods are shown. CERES provided the best results overall (MAGeT results were better for lobule I-IV, Vi and X although differences were not significant). In Figure 10, some examples from the different methods are shown. Note how SUIT tends to produce oversegmentation and always fills the fissures between lobules. Also note the irregular boundary between lobules I-IV (red) and V (yellow) produced by RASCAL in all three cases. Moreover, in the worst case, it can be seen that lobule VIIIA (purple) in MAGeT and RASCAL segmentations is barely visible. Finally, one of the main advantages of our method is its efficiency as it produces competitive results in terms of accuracy with a reduced computational time. The slowest method in the comparison was MAGeT (approximately 4 hours per case) due to the need of performing many nonlinear registrations. RASCAL takes around 110 minutes to provide the segmentation result. Finally, SUIT and CERES require around 5 minutes to produce the segmentation.

To conclude this section, it is important to note that CERES2 (CERES augmented with an error correction step) won the MICCAI 2017 international challenge ENIGMA on automatic cerebellum parcellation (Carass et al. 2018). During this challenge, nine methods were evaluated on 2 datasets, an adult dataset and a pediatric dataset. CERES2 significantly outperformed all the other methods including last Deep Learning (DL) framework. Table 4 presents a comparison on the pediatric dataset of CERES2 with the second and the third ranked methods, both based on DL strategies. The pediatric dataset is composed of 20 training images and 10 testing images. Mean DICE is provided across four hierarchies (Coarse, Lobe, Vermis, Lobule) of labeling and also the combination of all 28 labels (Consolidated). The method proposed in (Dolz et al. 2018) is a Fully Convolutional Network (FCN) based on 13 layers. DeepNet (Fonov and Collins 2018) is based on a U-Net architecture of 10 layers (Ronneberger et al. 2015). These results demonstrate that the proposed PBS method is still very competitive (significantly better at $p<0.0001$ for almost all the hierarchies) compared to the last DL frameworks.

*Table 4: Mean DICE coefficient values obtained during the ENIGMA challenge (Carass et al. 2018) on the pediatric dataset for FCN a fully convolutional network (Dolz et al. 2018), DeepNet based on U-Net architecture (Fonov and Collins 2018) and our CERES2 method (Romero et al. 2017a). Best results are marked in bold and significant differences ($p<0.0001$) are marked with * for FCN and ‡ for DeepNet comparison.*

Hierarchy	FCN (Dolz et al. 2018)	DeepNet (Fonov and Collins 2018)	CERES2 (Romero et al. 2017a)
Coarse	0.9326	0.8859	**0.9348** ‡
Lobe	0.8859	0.8827	**0.9033** *‡
Vermis	0.8491	0.8427	**0.8763** *‡
Lobule	0.8776	0.8808	**0.9043** *‡
Consolidated	0.8828	0.8815	**0.9043** *‡

1.4.3.2 Multimodal Hippocampal Subfields Segmentation: HIPS

In (Romero et al. 2016, Romero et al. 2017b), we introduced a new multimodal pipeline for hippocampal subfields segmentation called HIPS. The current hippocampal subfields definition is mainly based on the wotk of Lorente de No (1934). Starting from the upper end at the hippocampal sulcus, we find the dentate gyrus (DG) followed by the Cornu Ammonis (CA) which is subdivided in four consecutive parts (CA4 to CA1) and the Subiculum at the bottom end. The CA is also structured in six layers called stratum. These layers are the Stratum oriens (SO), Stratum pyramidale (SP), Stratum lucidum (SLU), Stratum radiatum (SR), Stratum lacunosum (SL) and the Stratum molecuare (SM). Currently, subfield segmentation protocols have been developed based on high resolution in-vivo MRI. In (Winterburn et al. 2013), the authors presented a new in-vivo high resolution atlas to divide the hippocampus in five different subregions: CA1, CA2-3, CA4/DG, Stratum and Subiculum as shown in Figure 11. Latter in (Kulaga-Yoskovitz et al. 2015), the authors developed another segmentation protocol consisting of three structures: CA1-3, CA4/DG and Subiculum as shown in Figure 11. However, there is still little consensus between the different HIPP subfield protocols as shown in (Yushkevich et al. 2015a) where 21 delineation protocols were compared.

The proposed HIPS method enables subfields segmentation according to the Winterburn and Kulaga-Yoskovitz protocols. This method uses a multimodal version of OPAL (Giraud et al. 2016) to produce fast and accurate segmentations in combination with nonlinear registration to build subject's specific library and involving a systematic error correction scheme. As explained at the end of the previous section, the use of this 3-stage framework is our current pipeline strategy to produce fast high quality segmentation. Due to its high resolution training libraries, HIPS can work with high resolution (0.5 mm^3) T1w and T2w images. However, during our validation, we showed that the proposed approach performs well on monomodal T1w and T2w and when using standard resolution images up-sampled using our nonlocal super resolution methods (Manjon et al. 2010b, Coupe et al. 2013). Finally, a new neural network-based error correction was proposed to minimize systematic segmentation errors at post-processing (see original paper for details). To validate our approach, we compared HIPS with other recent methods applied to hippocampus subfield segmentation on two datasets. For the Winterburn dataset, we compared HIPS with MAGeT (Pipitone et al. 2014) and for the Kulaga-Yoskovitz dataset we compared HIPS with ASHS a multi-template method including error correction (Yushkevich et al. 2015b) and SurfPatch combining surface-based and patch-based strategies (Caldairou et al. 2016).

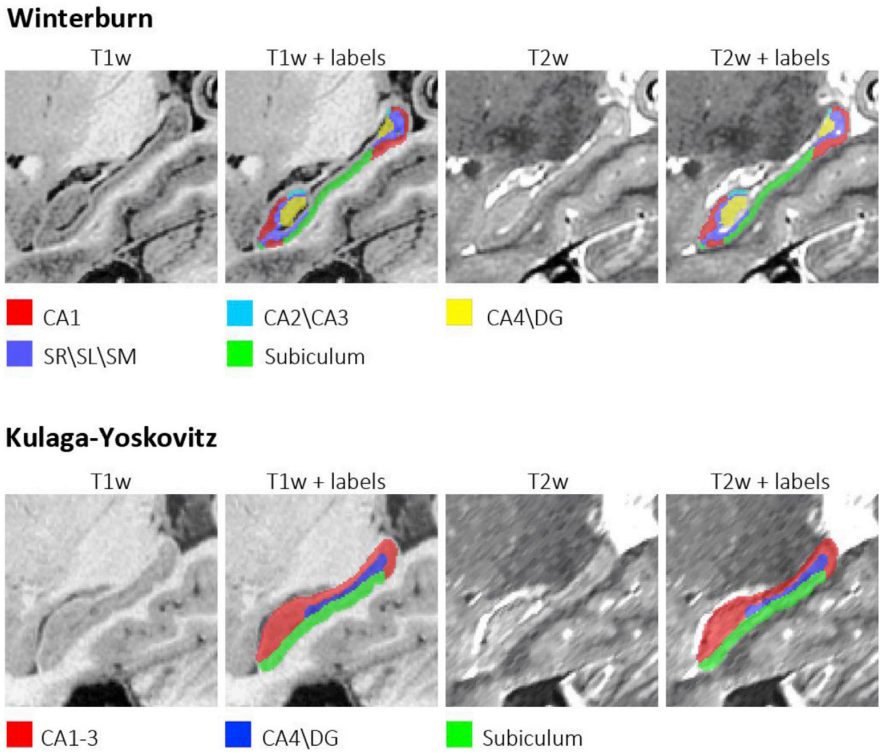

Figure 11: Examples from Winterburn and Kulaga-Yoskovitz datasets showing T1w, T2w and manual segmentations.

The Winterbrun dataset contains 5 subjects with high resolution T1w and T2w images and their corresponding manual segmentations. The high resolution images are publicly available at the CoBrALab website (http://cobralab.ca/atlases). These MR images were taken from 5 healthy volunteers (2 males, 3 females, aged 29–57). High resolution T1w and T2w images were acquired at an isotropic resolution of 0.6 mm^3. The hippocampi and each of their subfields were segmented manually by an expert rater including 5 labels (CA1, CA2/3, CA4/DG, (SR/SL/SM), and subiculum). For more details about the labeling protocol see the original paper (Winterburn et al. 2013). The Kulaga-Yoskovitz dataset includes 25 subjects from a public repository (http://www.nitrc.org/projects/mni-hisub25) (31 ± 7 years, 12 males, 13 females) with manually-drawn labels dividing the HIPP in three parts (CA1-3, DG-CA4 and Subiculum). MRI dataset consist of an isotropic T1w (0.6 mm^3) and anisotropic 2D T2w images (0.4 × 0.4 × 2 mm^3).

Table 5: Mean DICE in the native space for each structure. Segmentation performed by MAGeT (a multi-template method) and HIPS at 0.9mm^3 on the Winterburn dataset. Best results are in bold. We used the results provided in the native space from the corresponding publication.

Structure	MAGeT (Pipitone et al. 2014)	HIPS (Romero et al. 2016)
Average	0.5260	**0.6610**
CA1	0.5630	**0.6700**
CA2/CA3	0.4120	**0.5220**
CA4/DG	0.6470	**0.7630**
SR/SL/SM	0.4280	**0.5990**
Subiculum	0.5800	**0.7220**
Hippocampus	0.8160	**0.8760**

Table 6: Mean DICE in the native space for each structure. Segmentation performed by ASHS (Yushkevich et al. 2015b), SurfPatch (Caldairou et al. 2016), HIPS and human rater (intra-rater and inter-rater) on the Kulaga-Yoskovitz dataset. Best results (for automatic segmentation) are in bold.

Structure	ASHS	SurfPatch	HIPS	Inter-rater	Intra-rater
Average	0.8513	0.8503	**0.8744**	0.8833	0.9113
CA1-3	0.8736 ± 0.0197	0.8743 ± 0.0247	**0.9030 ± 0.0138**	0.8760 ± 0.048	0.9290 ± 0.010
CA4\DG	0.8254 ± 0.0345	0.8271 ± 0.0285	**0.8497 ± 0.0332**	0.9030 ± 0.036	0.9000 ± 0.019
Subiculum	0.8548 ± 0.0243	0.8495 ± 0.0245	**0.8705 ± 0.0212**	0.8710 ± 0.053	0.9050 ± 0.016

For the comparison, we used the results provided in the native space from the corresponding publication. Table 5 shows results for MAGeT and HIPS on the Winterburn dataset while Table 6 shows results for ASHS, SurfPatch and HIPS on the Kulaga-Yoskovitz dataset. For a fair comparison between considered methods, all the DICE coefficients for HIPS have been calculated using the segmentations in native space (using the corresponding inverse affine registration). We showed that HIPS outperforms other *state-of-the-art* methods in term of segmentation accuracy achieving an overall DICE of 0.661 for the Winterburn dataset while MAGeT (Pipitone et al. 2014) obtained a DICE of 0.5260, and an overall DICE of 0.8744 for Kulaga-Yoskovitz while ASHS (Yushkevich et al. 2015b) obtained 0.8513 and SurfPatch (Caldairou et al. 2016) obtained 0.8503. HIPS is also faster than the other methods taking an average execution time under 20 minutes compared to several hours required by both other methods. Finally, Figure 12 presents results obtained with HIPS on both considered datasets. We can see that results produced by HIPS are very similar to manual segmentations. Despite low contrast between subfield, HIPS is able to accurately estimate boundaries.

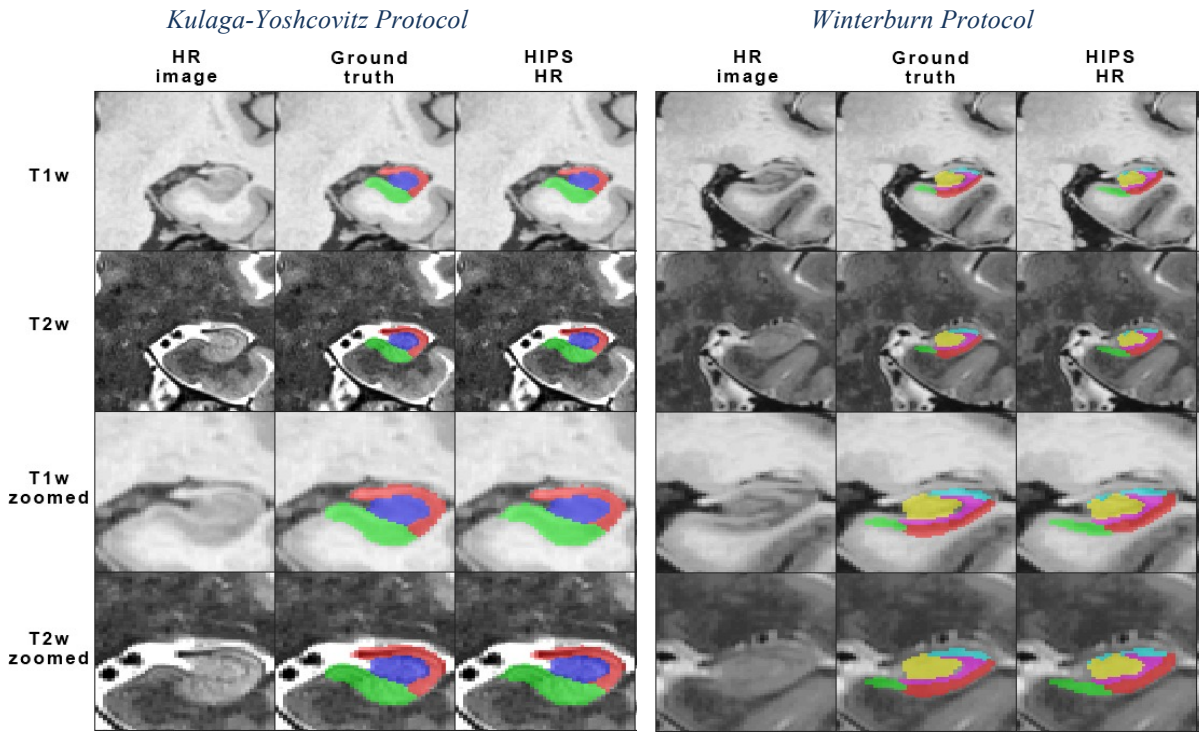

Figure 12: Results obtained with HIPS on the Winterburn and Kulaga-Yoskovitz datasets.

1.4.4 Multimodal Lesion Segmentation

In (Guizard et al. 2015a), we proposed a new multimodal PBS method to segment MS lesions. MS is a chronic, inflammatory demyelinating disease, which mainly affects the white matter of the central nervous system but may also affect the cortex. The disease presents itself with a wide range of clinical manifestations, usually beginning with a relapsing remitting (RRMS) phase. RRMS is characterized by attacks of worsening neurologic function (relapses) that are followed by partial or full recovery (remissions). Relapses are directly related to an underlying inflammation of the central nervous system, which affects the myelin of the axons and consequently leads to focal "MS lesions". Because MRI is sensitive to inflammatory and demyelinating changes, it is often used to monitor, identify and quantify MS lesions that are hyperintense on T2w MRI and may become hypointense on T1w images. Lesion counts are often used to assess the disease burden and track disease progression as new lesions are related to current disease activity. Both counts are used to assess the efficacy of new therapies. In the method proposed in (Guizard et al. 2015a), we focused on lesions commonly called "T2-lesions" (those that are hyperintense on T2w images) and did not consider other sub-types of lesions (i.e., gadolinium enhancing "active lesions", "black holes" and cortical lesions). MS lesions in MR images are extremely difficult to identify because of inter-subject anatomical variability, lesion location, size and shape. Manual segmentation of MS lesions is still recognized as the gold standard in MS, but it is time consuming and subjects to important intra- and inter-expert variability.

The method proposed in (Guizard et al. 2015a) uses a multimodal PBS to be able to segment MS lesions. During our experiment, we used T2w and FLAIR images only. Our method was validated using the dataset provided by the international MICCAI 2008 challenge (Styner et al. 2008). From the MS challenge website, 20 training MR datasets with ground truth manual lesion segmentations and 23 testing cases could be downloaded. While lesions masks for the 23 testing cases are not available for download, an automated system is available to evaluate the output of a given segmentation algorithm. At the time of paper writing in 2014, our segmentation results on the testing MICCAI challenge dataset were submitted online and compared with other published techniques including i) LesionTOADS (Shiee et al. 2010), ii) the winner of the MICCAI challenge (Souplet et al. 2008), iii) a supervised technique proposed in (Geremia et al. 2011) and iv) the method proposing the best score when our method was evaluated (Tomas-Fernandez and Warfield 2011). In this manuscript, I decided to update these results by adding the last available results based on DL methods – a Fully Convolutional Neural Network (FCN) presented in (Brosch et al. 2016) and the cascade of Convolutional Neural Network (CNN) presented in (Valverde et al. 2017).

Table 7: Results on the multimodal MICCAI Challenge dataset. Bold results were the best results at the time of the publication of our method (Guizard et al. 2015a). Cascade of CNN results (Valverde et al. 2017) have been added as the best results available at the time of habilitation thesis writing.

Method	Score
Lesion TOADS (Shiee et al. 2010)	79.96
EM classification (Souplet et al. 2008)	80.00
Random Forest (Geremia et al. 2011)	82.07
FCN (Brosch et al. 2016)	84.07
Outliers detection (Tomas-Fernandez and Warfield 2011)	84.46
Proposed PBS (Guizard et al. 2015a) (T2w + FLAIR)	**86.11**
Cascade of CNN (Valverde et al. 2017) (T1w + FLAIR)	86.70
Cascade of CNN (Valverde et al. 2017) (T1w + T2w + FLAIR)	**87.12**

The organizers combined different metrics to produce a score between 0 and 100, where 100 is a perfect score and 90 is the typical score of an independent rater (Styner et al. 2008). The MICCAI challenge website provides a results archive, allowing us to compare the performance of our method with other groups. The results are summarized in Table 7. At the time of paper writing, our method held the best result with an overall average summary score of 86.1 (note that 90 corresponds to a segmentation accuracy reaching human expert inter-rater variability). Compared to the FCN approach presented in (Brosch et al. 2016) – published one year after our work – our PBS is quite competitive (84.07 vs. 86.11). Finally, two years later, the results obtained with our PBS method using 2 modalities are still competitive compared to the last advanced DL method based on a cascade of CNN. Indeed, our method obtained 86.11 vs. 86.70 for cascade of CNN when using the same number of modalities, and 86.11 vs. 87.12 when cascade of CNN is trained on 3 modalities. Whereas many methods require at least 3 MRI contrasts (T1w, T2w, PDw or FLAIR) (Souplet et al. 2008, Geremia et al. 2011), and others require even-more contrasts (FLAIR, diffusion tensor imaging, fractional anisotropy and mean diffusivity, ...) (Morra et al. 2008), we use only two modalities (T2W and FLAIR). This dual-contrast method presents multiple advantages like decreasing the risk of corruption due to image artifacts, reducing the financial cost and increasing patient comfort. Figure 13 shows segmentation results obtained with our multimodal PBS method using T1w, T2w and FLAIR MRI on a RRMS case from a clinical study.

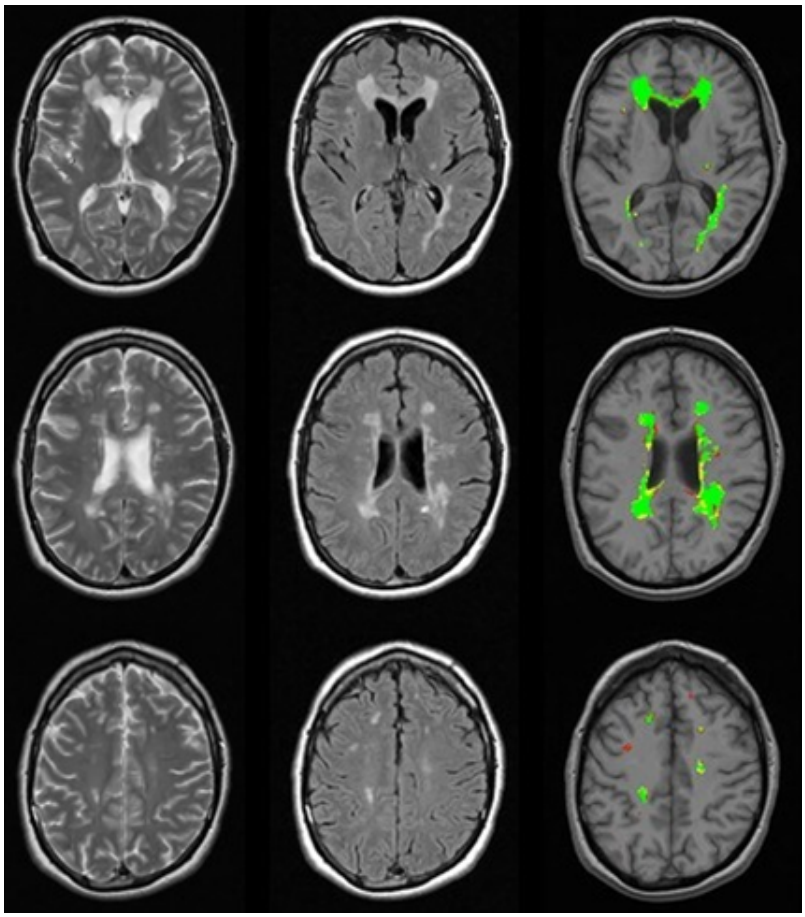

Figure 13: Segmentation results for one RRMS cases at different axial positions. The figure shows axial for T2w, FLAIR and T1w, the automatic segmentation obtained with our PBS method. The overlapping voxels (TP) with the manual segmentation are represented in green, while the false positives (FP) are yellow and the false negatives (FN) are red.

1.5 Discussion

In this first chapter, we presented our patch-based segmentation framework and some extensions that we proposed over the past years. After our pioneering works in (Coupe et al. 2010a, Coupe et al. 2011) for hippocampus segmentation, our PBS method has been the focus of many attention since the proposed method has demonstrated high accuracy in many applications despite its simplicity.

In (Eskildsen et al. 2012), we proposed a new brain extraction method, called BEaST, based on PBS segmentation embedded within a multiresolution framework. The accuracy of the method is higher than BET (Smith 2002), VBM8 and similar to a multi-template method based on multiple nonlinear registrations called MAPS (Leung et al. 2011), while being much faster and requiring a smaller library of expert priors. Using all baseline ADNI1 dataset, the study demonstrated that PBS segmentation is robust to pathology and consistent if the right priors are available. In (Manjon et al. 2014), we presented an improved version of BEaST for intracranial cavity extraction called NICE. The proposed improvements enabled to increase segmentation quality and reduce the computational load at the same time (the proposed method is able to work in approximately 4 minutes). NICE is now integrated in our online volBrain platform as described in Chapter 3.

In (Romero et al. 2015), we have presented an accurate and fast PBS segmentation method, called NABS, for segmenting cerebral and cerebellar hemispheres and brainstem. We did not present NABS in this chapter but the main methodological novelties over other similar methods are the use of a multi-label block-wise label fusion strategy specifically designed to deal with the classification of large brain structure. We have shown that NABS method was able to accurately segment brain structures in healthy subjects across a wide range of ages. As explained later in Chapter 3, we have also provided quantitative comparisons against the ADisc method (Zhao et al. 2010), which represents the *state-of-the-art* for brain hemisphere segmentation. This comparison demonstrated the competitive performance of NABS. NABS is also included in our online volBrain platform.

In (Guizard et al. 2015a), we proposed a new method to detect MS lesions using a training library containing T2w and FLAIR images along with manual T2w lesion masks. This adaptation of our PBS segmentation method to MS lesions identification with a new multi-contrast measure has proven to be highly competitive in our internal validation and in an independent comparison. Our method provides segmentation quality near inter-rater variability for MS lesion segmentation. This remains competitive compared to last deep learning methods (Brosch et al. 2016, Valverde et al. 2017). Finally, we integrated it into our pipeline called lesionBrain freely available via our onle volBrain platform.

In (Ta et al. 2014, Giraud et al. 2016), we proposed a novel PBS method based on an optimized PatchMatch label fusion. Thanks to the low computational burden of this method, we investigated the potential of a new multi-feature and multi-scale framework with late estimator aggregation. The validation of our approach on hippocampus segmentation applied to two different datasets shows that the proposed method produces competitive results compared to *state-of-the-art* approaches. Indeed, OPAL obtained the highest median Dice coefficient with a drastically reduced computation time. In addition, OPAL reaches the inter-expert reliability on both datasets. Therefore, OPAL provides automatic segmentations equivalent in terms of Dice coefficient to inter-expert segmentations in less than 2s of processing for the segmentation step.

In (Romero et al. 2017a), OPAL has been extended to perform cerebellum lobules segmentation. We have presented a new pipeline called CERES that works in a fully automatic manner and that is able to provide accurate results in a low computation time. We showed that CERES produces *state-of-the-art* results outperforming current cerebellum segmentation methods such as SUIT an atlas-based method (Patenaude et al. 2011), MAGeT a multi-template method (Park et al. 2014) and RASCAL a PBS method (Weier et al. 2014). In terms of accuracy, CERES obtained the best results for most of the labels. More interestingly, CERES almost reaches intra-rater accuracy for the whole cerebellum segmentation (CERES = 0.9377 vs. human = 0.941). From the efficiency point of view, CERES was found to produce the best segmentation results in a very short time (5 minutes per case). Moreover, CERES2 (including the error correction step proposed for HIPS) won the international MICCAI challenge ENIGMA in 2017 with significantly better performance than several fully convolutional network methods (Carass et al. 2018). Finally, we have integrated CERES as a part of our online volBrain platform.

Similarly, in (Romero et al. 2016, Romero et al. 2017b), we applied the CERES framework to hippocampus subfields segmentation by extending it to multimodal case. Moreover, we added a correction error step based on machine learning in a way similar to (Wang et al. 2011). The proposed framework HIPS yielded to very competitive approach. As previously mentioned, we are now using a 3-stage framework: 1) construction of a subject specific training library by using precomputed non-linear registration, 2) fast PBS using OPAL and 3) error correction based on machine learning. We are currently integrating HIPS in our online volBrain platform.

In addition to the extensions that we proposed, several groups proposed key contributions to improve PBS framework. In (Wang et al. 2013a), the authors proposed a new voting scheme to improve segmentation accuracy. In (Asman and Landman 2013), the authors developed a novel statistical label fusion algorithm based on an expectation-maximization algorithm. This enables to reduce the need for large atlas sets. In (Bai et al. 2013), the authors proposed a Bayesian framework to better describe PBS. Similarly, in (Wu et al. 2014), the authors proposed a generative probabilistic model in order to enforce the labeling consistency. In (Wang et al. 2014b), the authors adapted PBS to infant brain segmentation that is a challenging problem due to low contrast. Similarly, in (Liu et al. 2016), PBS is adapted to segment brain of premature neonates. In (Cordier et al. 2016a), PBS is successfully extended to tumor segmentation. In (Wu et al. 2015b), the authors proposed to hierarchically improve the label fusion accuracy by dynamically changing the patch size to improve segmentation accuracy. In (Bai et al. 2015), the authors proposed to use augmented features such as gradient and contextual information. Similar idea has been applied in (Wachinger et al. 2017a) using a large range of local descriptors. Finally, in (Cordier et al. 2016b), patch-based framework has been used to perform multimodal image synthesis of pathological cases.

Moreover, our PBS strategy has been applied in variety of clinical settings. We used PBS to study Parkinson's disease (Haegelen et al. 2013), multiple sclerosis (Moroso et al. 2017, Planche et al. 2017), glaucoma (Tellouck et al. 2016), schizophrenia (Huhtaniska et al. 2017) and Alzheimer's disease (see Chapter 2). Other groups extended our PBS to other modalities such as CT (Liao et al. 2013, Wang et al. 2014a) or ultrasound (Yang et al. 2015), but also to other organs such as heart (Bai et al. 2013, Bai et al. 2015, Zhuang and Shen 2016), prostate (Liao et al. 2013) or knee (Wang et al. 2013b, Shan et al. 2014). Finally, PBS method has been successfully applied to the problem of abdominal multi-organs segmentation (Wolz et al. 2013, Tong et al. 2015). For each of these applications, PBS demonstrated high robustness and efficiency compared to *state-of-the-art* methods.

Nowadays, PBS strategy is considered as a *state-of-the-art* method and remains competitive compared to last advanced machine learning methods such as DL for anatomical structure segmentation. In fact, despite the great success of DL in computer vision, recent brain MRI segmentation methods based on CNN (Wachinger et al. 2017b) did not outperform PBS methods proposed several years ago. For instance, whole brain segmentation based on CNN (Wachinger et al. 2017b) did not perform better than PBS proposed in (Wang and Yushkevich 2013). Moreover, as previously discussed, fully convolutional networks (Dolz et al. 2018, Fonov and Collins 2018) performed worse than CERES2 during the MICCAI 2017 international challenge ENIGMA (Carass et al. 2018).

So far, one of the main issues of using DL in medical imaging is *" the limited number of training scans with manual segmentations"* (Wachinger et al. 2017b) while DL approaches required a huge number of training examples. One way to overcome this limitation is to use patch-wise DL. By splitting one training image in thousands of patches, the training library size is drastically increased. This is well-adapted to MS lesion segmentation since multiple occurrences of the target object appear in a single image. Consequently, this explains why patch-wise cascade of CNN presented in (Valverde et al. 2017) is currently one of the best method for this application and outperforms our PBS. It has to be noted that there is a growing trend to use patchwise CNN (Ghafoorian et al. 2017, Guerrero et al. 2018) or FCN (Xu et al. 2017, Li et al. 2018) for white matter hyper intensities. This is assessed by the large majority of DL-based methods proposed during the last MS MICCAI challenge[1].

For anatomical structure segmentation, where few occurrences of the same object are present in a single image, such DL strategy is less efficient which explains the results obtained in (Wachinger et al. 2017b, Dolz et al. 2018, Fonov and Collins 2018). Nevertheless, as we studied in (Eskildsen et al. 2012, Giraud et al. 2016), one way to increase the training library size is to propagate manual segmentations over a large number of unlabeled scans by using automatic segmentation methods (Wolz et al. 2010). Indeed, as presented in Chapter 3, it is now possible to get access to a large number of non-labeled MRI scans thanks to open access databases. We plan to investigate this strategy in the future in order to address one of the current limitations of using DL for medical images.

[1] https://portal.fli-iam.irisa.fr/msseg-challenge/overview

Chapter 2 Computer-Aided Diagnosis

Abstract: In this second chapter, we will present the adaptation of our patch-based segmentation framework to patch-based grading (PBG) of anatomical structures. The grading of the structure under consideration is achieved by estimating the nonlocal similarity of the subject to different training populations. Our PBG method estimates at each voxel a score reflecting the degree of the pathology and thus enables to perform computer-aided diagnosis. After a general principle description, we will detail different extensions such as ensemble-based grading, multimodal grading or multi-feature strategy. Then, we will apply our PBG method to hippocampal grading and whole gray matter grading for Alzheimer's disease diagnosis. Different experiments will be proposed to validate the developed computer-aided diagnosis tools. We will show that our PBG framework and its extensions have competitive performance compared to recent state-of-the-art methods. Moreover, we will investigate the interest of performing multimodal hippocampal subfields grading. Finally, we will present the prognosis performance of our method at presymptomatic phase using a long follow-up dataset.

2.1 Introduction

Over the past decade, computer-aided diagnosis (CAD) was a rapidly growing field of research. This field aims at developing new image analysis techniques to assist the clinicians to interpret images. Usually based on automatic pipeline involving preprocessing, feature extraction and feature classification; CAD can be dedicated to pathology detection or to predict its evolution. For some tasks, such as the simultaneous comparison of a large number of images or the detection of subtle anatomical changes caused by diseases, computer is now an essential tool.

Our works dedicated to CAD over the past five years were mainly focused on Alzheimer's disease (AD) detection and prediction (Coupe et al. 2012a, Coupe et al. 2012b, Eskildsen et al. 2013, Komlagan et al. 2014, Coupé et al. 2015, Eskildsen et al. 2015, Hett et al. 2016). AD is the most common form of dementia affecting the elderly and the prevalence of AD increases with age. Moreover, it is the fourth leading cause of death among adults in high-income countries. Although numerous drug-modifying clinical trials for AD have been conducted, so far none has been effective (Karran and Hardy 2014). Two hypotheses could explain this lack of efficiency:

- First, therapy was implemented too late after irreversible brain damage occurred (Cummings et al. 2007, Callaway 2012). In fact, when cognitive function alterations are important enough to be used for diagnosis, the pathological burden is already high and therefore the brain damages are pronounced (see Figure 14 on the left). This highlights the need to identify the disease earlier.
- Second, the therapeutic strategy is not appropriate and thus this requires a better understanding of disease pathological mechanisms.

In both cases, finding very early biomarkers of prodromal AD, characteristic of the presymptomatic phase (before memory loss and cognitive decline) of the disease, is therefore crucial. The development of such biomarkers can make easier the design of clinical trials and thus accelerate the development of new therapies.

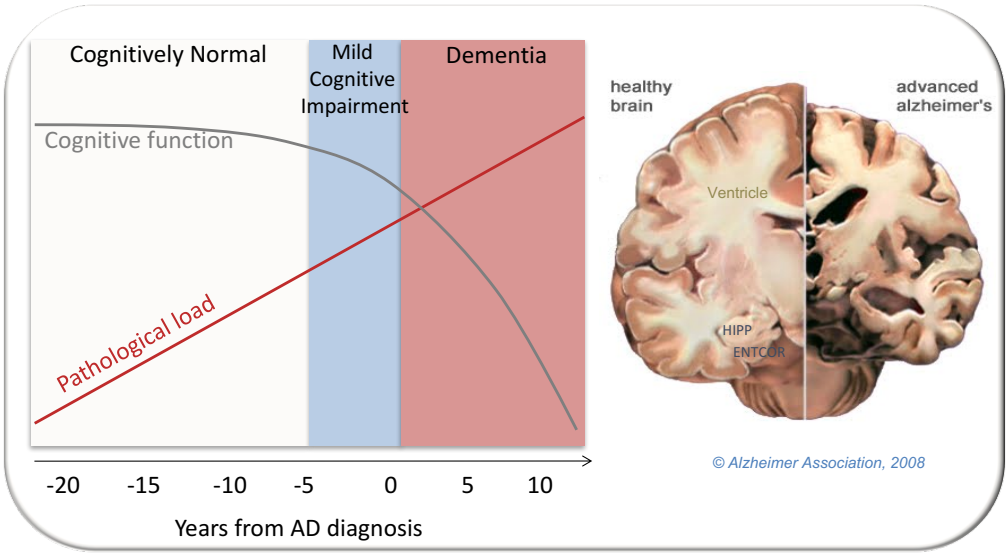

Figure 14: Left: the different stages of pathology progression, from presymptomatic phase of the cognitively normal subject to dementia of the patient with AD. Right: typical brain alterations caused by AD (e.g., atrophy of hippocampus and entorhinal cortex, and lateral ventricles enlargement).

Research from diverse scientific disciplines has focused on the identification of the earliest prodromal signs and risk factors for AD (Ballard et al. 2011). Many potential biomarkers have already been studied in depth with the goal of achieving this task (Jack et al. 2013). For example, the presence of amyloid -β, a hallmark of AD, seems to occur in the very early course of the pathology, long before the typical clinical, behavioral, and social criteria of dementia are fully met (Frisoni et al. 2010). Amyloid-β presence can be studied using cerebrospinal fluid (CSF) markers or positron emission tomography (PET). The possibility to directly detect amyloid during life is very interesting in establishing an early diagnosis of AD – see Figure 15, where amyloid-based markers are the earliest ones). However, so far, the results found are heterogeneous, and therefore, the links between amyloid-β burden and cognitive deficits are still unknown (Aizenstein et al. 2008, Chetelat et al. 2010, Villemagne et al. 2011, Kantarci et al. 2012, Lehmann et al. 2013, Jung et al. 2016). Biomarkers based on MRI are also increasingly under investigation because they are considered sensitive to the progression of AD at the predementia stage and correlated with cognitive decline (Frisoni et al. 2010). Usually, these imaging biomarkers are used to detect abnormal patterns of atrophy caused by AD on key structures in the brain.

The structures in the medial temporal lobe (MTL) are being studied because of their strong involvement in the pathogenesis of AD (Braak and Braak 1991). Recent MRI studies have also contributed to a better understanding the structural changes underlying AD cognitive impairment by demonstrating the association of cognitive difficulties with reductions in hippocampal volume (de Jong et al. 2008). Accordingly, histopathological investigations (Braak and Braak 1991) suggest that AD begins with the formation of neurofibrillary tangles in the MTL, particularly the entorhinal cortex (ENTCOR), a structure of the parahippocampal cortex. This formation then continues in the hippocampus (HIPP) and from there expands to other structures throughout the neocortex (see Figure 14 on the right). Therefore, using MTL structure atrophy as early imaging biomarkers is considered a promising way to follow the progression of AD (Frisoni et al. 2010), especially since changes in these structures are closely related to modifications in the subject's cognitive performance.

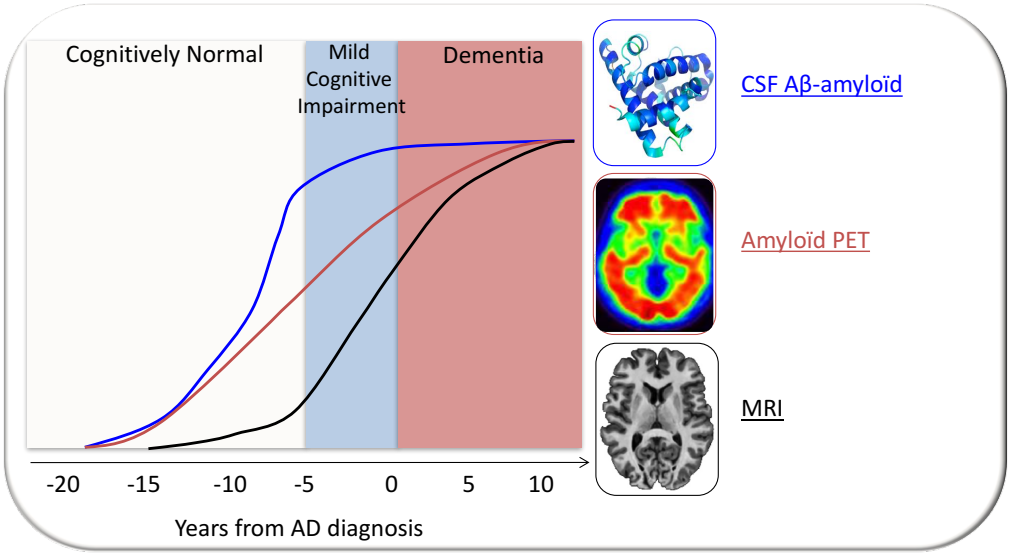

Figure 15: Hypothetical model of Alzheimer's disease biomarkers proposed by Jack et al. (2013). Progression of different biomarkers during the evolution of the pathology.

However, the automatic extraction of these MTL structures is challenging, especially in the case of the ENTCOR (Du et al. 2001). Moreover, the inter-subject variability of brain anatomy tends to limit the performance of AD detection methods that use only volumetric approaches (Wolz et al. 2011b, Coupe et al. 2012a). These two aspects limit the ability of volume-based imaging biomarkers that use MTL structures to characterize the earliest stages of AD as well as to develop efficient prevention or early intervention strategies.

To overcome limitations of volume-based imaging biomarkers, we proposed a new patch-based grading (PBG) method providing a better characterization of structure atrophy (Coupe et al. 2012a, Coupe et al. 2012b). Our PBG method estimates at each voxel a score reflecting the degree of the pathology. The grading of the structure under consideration is achieved by estimating the nonlocal similarity of the subject to different training populations (see Figure 16). Because it uses a nonlocal framework, our PBG framework addresses the problem of inter-subject variability nicely by enabling a one-to-many mapping between the subject's anatomy and those of the training templates. Moreover, by employing the patch-based comparison principle, our PBG is able to detect subtle changes caused by the disease.

In this chapter, we will introduce first the principle of our PBG method. Afterwards, we will present its application to hippocampus as well as its extension to whole gray matter (GM) grading (Komlagan et al. 2014) and to multimodal grading based on MRI and diffusion MRI (Hett et al. 2016, Hett et al. 2018). Moreover, we will discuss the classification results obtained with our PBG for different stage of the pathology. First, we will present AD diagnosis performance of our PBG compared to usual biomarkers. Second, we will show AD prognosis accuracy of our PBG at early stage of AD (i.e., on subjects with MCI few years before AD diagnosis). To this end, we will compare its performance with recent DL methods. Finally, we will study the ability of our PBG to detect subjects with high risk to develop AD at the presymptomatic stage (i.e., on cognitively normal subjects without clinical symptoms who will develop AD several years later).

2.2 Methods

2.2.1 From Segmentation to Grading

In (Coupe et al. 2012a, Coupe et al. 2012b), we proposed to extend our PBS method to efficiently aggregate clinical status – such as Cognitively Normal (CN) or AD – to estimate the proximity in the nonlocal means sense of each voxel compared to both populations constituting the training library (see Figure 16). To achieve this goal, we introduce the new concept of PBG that reflects the similarity of the patch surrounding the voxel under study with all the patches present in both training populations. In this way, the neighborhood information is used to robustly drive the search of anatomical patterns that are specific to a given subset of the training library. When the training populations include data from subsets of subjects in different clinical states, this approach provides an estimation of the grade (i.e., degree of closeness to one group or another) for each voxel:

$$g(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j}) p_s}{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j})} \quad \text{(Eq. 8)}$$

where p_s is the clinical status of the training subject s. In our case, $p_s=-1$ was used for AD status and $p_s=1$ for CN status. A negative grading value (respectively, a positive grading value) $g(x_i)$ indicates that the neighborhood surrounding x_i is more characteristic of AD than CN (respectively, of CN than AD). The absolute value $|g(x_i)|$ provides the confidence given to the grade estimation. When $|g(x_i)|$ is close to zero, the method indicates that the patch under study is similarly present in both populations and thus is not specific to one of the compared populations and provides little discriminatory information. When $|g(x_i)|$ is close to 1, the method detects a high proximity of the patch under study with the patches present in one of the training populations and not in the other. It has to be noted that other variables can be used in place of p_s such as patient's age to perform automatic age estimation or clinical scores to perform score prediction. In (Coupé et al. 2015), we showed that patch-based estimation of the Mini Mental State Evaluation (MMSE) (Folstein et al. 1975) provided more relevant information than binary pathological status p_s.

Finally, for each subject, an average grading value is computed over all voxels in the estimated segmentation of the structure of interest. Since the grading and the segmentation involve the same patch comparison step, the structures are extracted at the same time that their grade is estimated. Once all the subjects are processed using our PBG, the final step consisted in subject classification based on the extracted features (see Figure 17). During our works on CAD, we used different classifiers such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) or Random Forest.

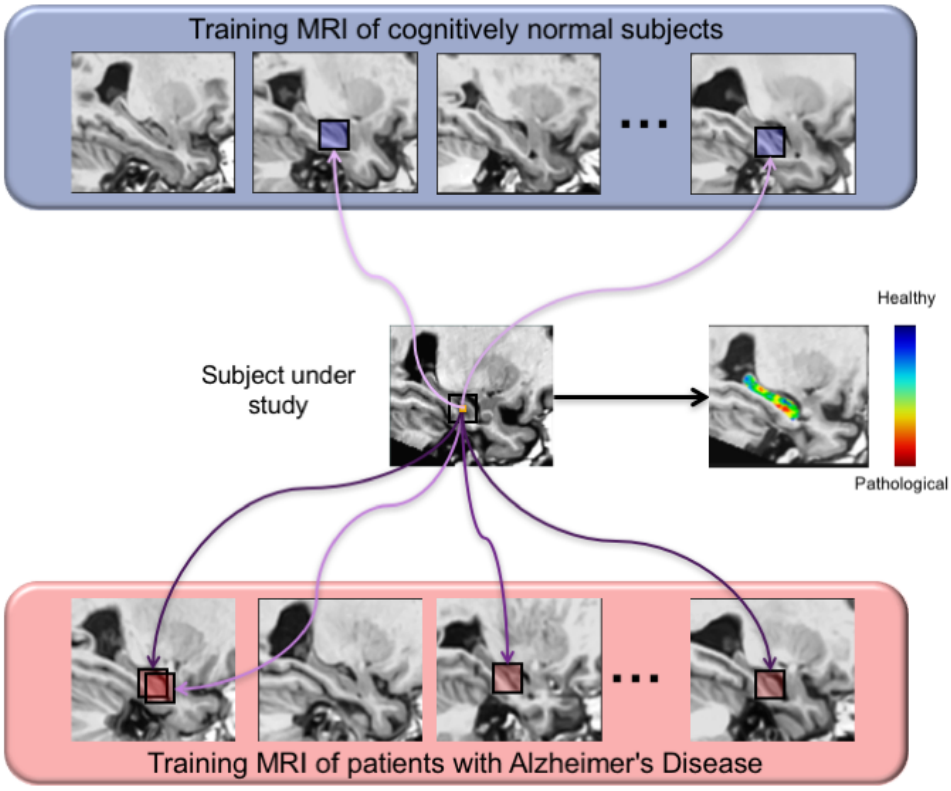

Figure 16: Global overview of the proposed patch-based grading. For each patch of the subject under study a comparison is performed with all the patches in the training subjects of both populations CN and AD (some examples of similar patches are displayed). Purple arrows symbolize the similarity between anatomical patterns. Dark purple indicates high anatomical patterns similarity and light purple indicates low similarity. For the anatomical pattern under study in this example more similar anatomical patterns are found in AD population and their weights (similarities) are higher. In the grading map, when the voxel color is blue, this indicates that the anatomical pattern surrounding this voxel is more similar to anatomical patterns found in the CN training subjects and thus that the local anatomy is closer to anatomies of healthy subjects (i.e. CN-like anatomical pattern). When the voxel color is green, this indicates that the anatomical pattern is equally similar to anatomical pattern found in the CN and the AD training subjects. Finally, when the voxel color is red, this indicates that the local anatomy contains alterations typical of the AD population (i.e. AD-like anatomical pattern) and thus the presence of neurodegeneration.

The grading concept relies on the same assumptions used by the nonlocal PBS, but it also performs an additive one. In fact, in grading, we consider that, in average, the patches extracted from an MRI of a new patient with AD will be more similar (in the nonlocal means sense) to patches extracted from training MRI of patients with AD than to patches extracted from training MRI of healthy subjects. To be usable, this approach has to be applied on a disease that has an impact on the patient's anatomy that is somehow detectable in MRI, where MRI needs only be sensitive to the change (MRI does not need to be specific). In addition, the grading study can be limited to key structures we know to be impacted by the particular disease. As a consequence, in our first studies on AD, we decided to apply grading on ENTCOR and HIPP two structures impacted early in the disease progression (Coupe et al. 2012a, Coupe et al. 2012b). However, this concept can be applied to other structures and modalities, and can be used to detect other diseases. In (Komlagan et al. 2014), we proposed to perform PBG over the whole GM. In (Hett et al. 2016), we applied our PBG to diffusion MRI. Finally, in (Hett et al. 2018), we proposed a multimodal PBG method based on MRI and diffusion MRI to better characterize HIPP subfields alterations. These extensions will be described in the next section.

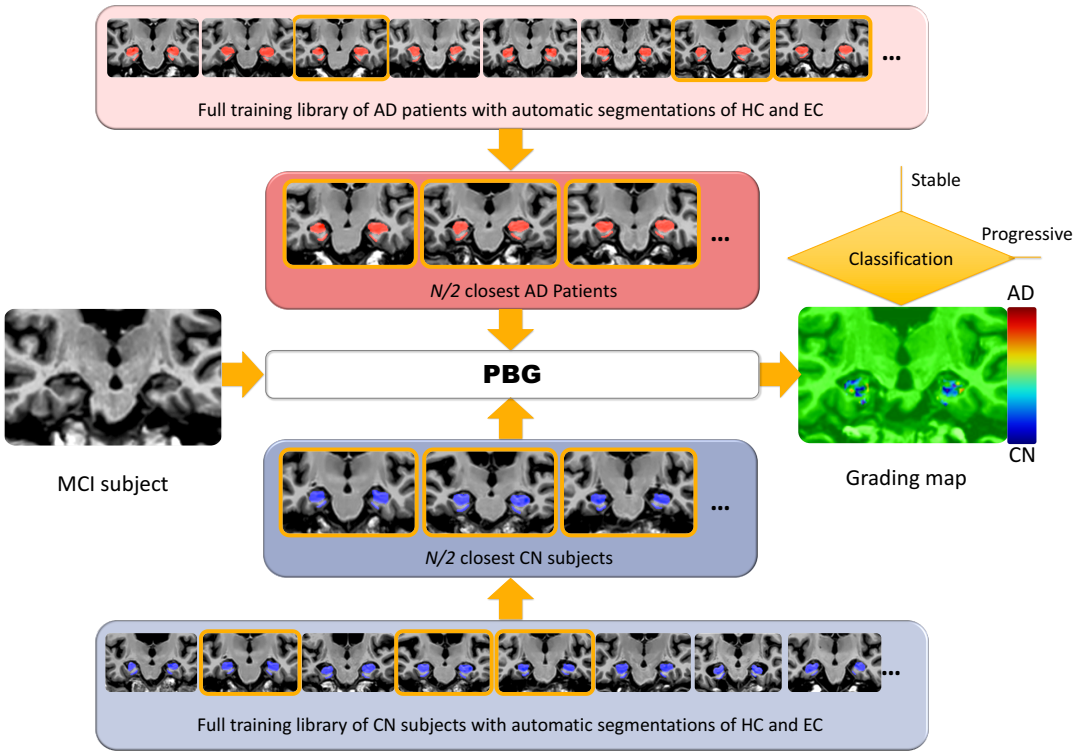

Figure 17: Example of our PBG prognosis workflow. First, PBG is used to estimate the grading maps of the test subject (an MCI subject in this example) using the training AD and CN subjects. The grade is estimated over the considered structure of interest, the HIPP in this example. Then, a classifier is trained using the HIPP grading values obtained on training AD and CN subjects. Finally, the test MCI subject is classified to obtain the final prognosis (i.e., stable or progressive).

2.3 Extensions

2.3.1 Whole Gray Matter Grading

In (Komlagan et al. 2014), we proposed to extend our PBG to the entire gray matter (GM). Indeed, the *a priori* definition of a region of interest may discard other possible informative anatomical regions. The choice between *i)* the noise reduction by using restricted areas that we know to be impacted by AD (knowledge-driven strategy) and *ii)* the use of all available information (data-driven strategy) is still an open question. Therefore, to investigate data-driven strategy, in (Komlagan et al. 2014), we proposed to grade the whole GM and to automatically perform structure selection. Moreover, to be more robust to intensity normalization discrepancies between MRI, probabilities were used in place of intensities during patch comparison. Finally, we presented an ensemble learning method to efficiently fuse the obtained grading values. Since the scoring/grading value, assigned to each voxel of the GM, estimates the proximity to AD and CN, it can be viewed as the posterior probability of a weak classifier. Combined together, these weak classifiers form an ensemble that can be used to classify subjects (Liu et al. 2012). As noticed in (Frisoni et al. 2010), it appears that AD-related brain alterations are mainly a region-by-region process. Hence, we proposed to create sub-ensembles of these weak classifiers using an atlas. Each of these sub-ensembles corresponds to an anatomical structure. At the end, to discard brain areas that may not be related to AD, we proposed to select the most relevant anatomical sub-ensembles using a Sparse Logistic Regression (SLR). The framework of the proposed method is summarized in Figure 18 and detailed in the following.

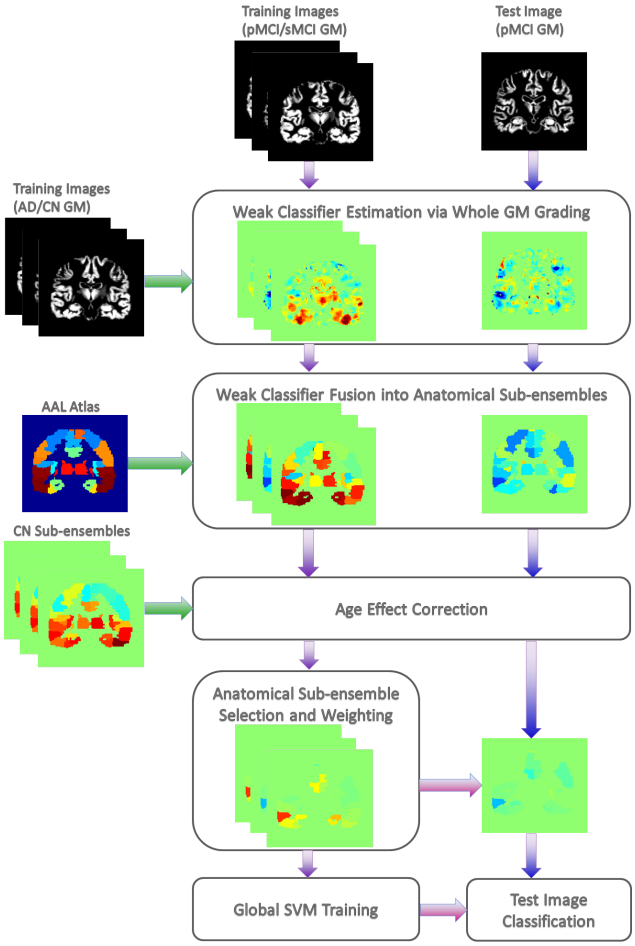

Figure 18: Anatomically constrained weak classifier fusion for whole GM grading. First, whole GM grading is estimated for all MCI subjects (progressive MCI and stable MCI) using CN and AD training library. Then, the Automatic Anatomical Labeling atlas is used to fuse local weak classifiers into anatomical sub-ensembles. For each sub-ensemble, the effect of age is corrected using the training CN. Afterwards, the training MCI images are used to perform sub-ensemble selection using a SLR. Finally, SVM classifier is trained using the selected sub-ensembles. At the end, the test MCI image is classified to obtained the final prognosis.

First, the grading method is applied to all the MCI subjects – progressive MCI (pMCI) and stable MCI (sMCI) – to obtain GM maps. In the proposed method, the grade of each GM voxel is calculated using probability of GM tissue instead of voxel intensities too be more robust to multi-site MR image acquisition. Second, the grading values obtained over the whole GM are fused into anatomical sub-ensembles to form intermediate classifiers. After the grading step, the dimensionality of the weak classifiers space is too high to be directly used for classification. A straightforward solution is to fuse all the weak classifiers into a global classifier (Liu et al. 2012). However, this may lead to a sub-optimal result since local relevant information may be lost in a high level global fusion. Additionally, as noticed in (Frisoni et al. 2010), AD affects specific regions of the brain in a typical progressive manner. Therefore, we proposed to group the weak classifiers into anatomical sub-ensembles using an atlas-based strategy. The sub-ensembles were constructed by an un-weighted vote of the weak classifiers included in each anatomical area. In (Komlagan et al. 2014), the whole GM was divided into the 116 segmented anatomical regions corresponding to the Automatic Anatomical Labeling atlas (Tzourio-Mazoyer et al. 2002). Thus, the grades are averaged within each anatomical structure and their mean values considered as grading values of the 116 anatomical structures.

Third, SLR feature selection was applied to select and weight the most relevant intermediate classifiers. As shown in (Jack et al. 2013), anatomical regions may not be similarly impacted by the progression from MCI stage to the moderate stage of AD. Therefore, using all the intermediate classifiers could be suboptimal since none impacted structures could be included. Moreover, beyond classification efficiency reasons and for clinical considerations, it could also be interesting to know the most impacted brain regions. In (Komlagan et al. 2014), we automatically selected the most relevant anatomical sub-ensembles by using SLR. During our experiments, the structures selected with the highest weights were the middle temporal lobe, the left hippocampus and the middle cingulum. Finally, the selected intermediate classifiers were used to train a global linear SVM classifier.

2.3.2 Multimodal Grading of Hippocampal Subfields

PBG method presented in previous sections was firstly designed to capture structural alterations in T1w MRI. Although anatomical MRI is a valuable imaging technique to measure structural modifications, such modality is not able to capture microstructural degradation. The microstructural modifications caused by AD are considered to occur before the atrophy measured by anatomical MRI. Therefore, diffusion MRI appears as a good potential candidate to detect the earliest sign of AD. Consequently, we recently proposed to apply our PBG on Diffusion Tensor Imaging (DTI) to detect microstructural modifications (Hett et al., 2016). In a first study, we showed the efficiency of HIPP grading based on mean diffusivity (MD) to improve the classification of the early stages of AD. Afterwards, we investigated the use of multimodal PBG (MPBG) on hippocampus subfields (Hett et al. 2018).

As mentioned in Chapter 1, the hippocampus has been one of the most studied structures for automatic detection of AD. However, this structure is complex and not homogeneous. HIPP is subdivided into several subfields each one having distinct characteristics. Several MRI studies demonstrated that subfields are impacted differently according to AD stages (Apostolova et al. 2006, Kerchner et al. 2010, Kerchner et al. 2012, La Joie et al. 2013). These results indicate that analysis of HIPP alterations at finer scale could provide better tool to study AD progression. Therefore, in (Hett et al. 2018), we proposed an innovative MPBG framework to fuse PBG extracted from different MRI modalities (i.e., T1w MRI and diffusion MRI) and we decided to apply this MPBG framework over hippocampal subfields obtained with HIPS (see Figure 19). This multimodal fusion strategy is detailed in the following.

First, for each modality a training library is built with CN and AD subjects. Then, a grading map is estimated for each considered modality. A straightforward strategy would be to average the obtained grading maps. However, the quality of the grading estimation is not the same for all the modalities all the locations. Thus, the grading value for a modality should be weighted according to the confidence of the local grading value. In (Hett et al. 2018), we proposed an novel framework to fuse several grading maps obtained from M different modalities. Our fusion strategy is based on the fact that estimated grading maps from different modalities may not have the same relevance, but more importantly all local weak classifiers in these maps do not have the same quality. Hence, at each location, we propose to combine weak classifiers derived from each modality according to a confidence criterion. Therefore, the grading value of a grading map of the modality m, denoted g_m, at voxel x_i, is weighted by $\alpha_m(x_i) = \sum_{x_{s,j}} w_m(x_i, x_{s,j})$ that reflects the confidence of g_m as shown in (Sutour et al. 2014). Thus, each grading map provides a weak classifier at each voxel location that is weighted with its degree of confidence $\alpha_m(x_i)$.

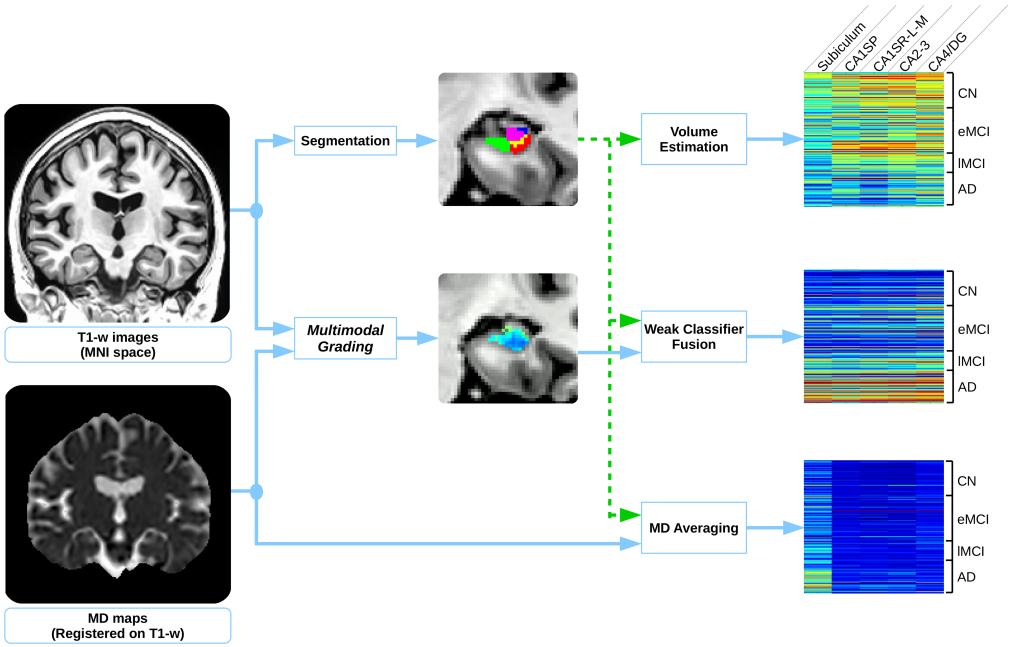

Figure 19: Multimodal PBG on hippocampal subfields. At left, the input data with both modalities: T1w images registered into the MNI space and MD maps registered to the T1w images. At the middle and from top to bottom: hippocampal subfields segmentation on T1w image and multimodal grading are computed. At right, features of the considered subfield biomarkers for the 4 studied groups AD, early MCI, late MCI and AD. From top to bottom, the features are the volumes, the multimodal grading values and the MD values.

At the end, the final grading value g_M resulting from our adaptive fusion strategy is given by:

$$g_M(x_i) = \frac{\sum_{m \in \mathrm{M}} \alpha_m(x_i) g_m(x_i)}{\sum_{m \in \mathrm{M}} \alpha_m(x_i)} \qquad \text{(Eq. 9)}$$

The proposed fusion framework is spatially adaptive and take advantage of having access to a local degree of confidence $\alpha_m(x_i)$ for each grading map m. In our previous works on multimodal segmentation (Guizard et al. 2015a), we used multimodal weight w_M (see Eq. 6) that estimates similarity over all the modalities at the same time. Here, the grading maps are estimated independently for each modality and fused *a posteriori* according to a local confidence criterion.

2.3.3 Multifeature Grading

As details in previous the chapter on PBS, we demonstrated that using edge detection filters can improve patch-based segmentation (Giraud et al. 2016). This result highlights that patches comparison can be improved by estimating patterns similarity on derivative features. Moreover, it is has been recently showed that HIPP texture plays a crucial role for the detection of early stages of AD (Sørensen et al. 2017). Therefore, in (Hett et al. 2017), we proposed to perform patch-based grading on multiple texture maps obtained with Gabor filters. Gabor filters were designed to detect salient features at specific resolution and direction. These filters were widely used for texture classification (Manjunath and Ma 1996). Consequently, the proposed strategy aims at the same time to improve the comparison of patches and to capture HIPP texture modifications (see Figure 20). This method is detailed in the following.

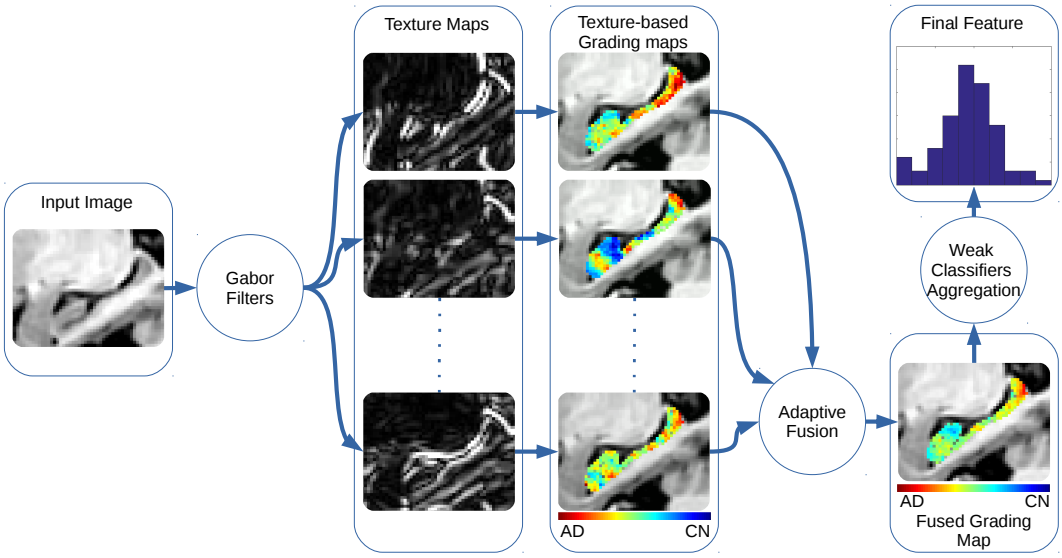

Figure 20: Proposed multifeature PBG framework.: from left to right, the T1w input data, the texture maps for different directions of Gabor filters, the intermediate texture-based grading maps for each direction, the final fused grading map and the histogram of weak classifiers.

First, the texture maps are estimated using 3D Gabor filters. In the proposed pipeline, the preprocessed MRI of the subject under study is filtered with a bank of Gabor filters to obtain multiple texture maps. It has to be noted that all the training library is also filtered with the same filters bank. Second, we applied our fast PBG based on OPAL on each texture maps to obtained multiple texture-based grading maps. Afterwards, we used the locally adaptive fusion scheme presented in previous subsection for MPBG (see Eq. 9) to fuse the multiple texture-based grading maps. Finally, we performed weak classifiers aggregation. In our initial works (Coupe et al. 2012a, Coupe et al. 2012b, Komlagan et al. 2014), the weak classifiers aggregation was performed using a simple averaging as explained previously. While using a strategy based on averaging enables to be robust to noise, this may remove relevant information on weak classifiers distribution. In (Hett et al. 2017), we proposed to approximate weak classifiers distributions using histograms. Consequently, we classified histogram bins with a SVM instead of classifying mean grading value over the segmentation mask (see Figure 20).

2.4 Results

In this section, we will present the main results obtained by our PBG for automatic AD diagnosis (Coupe et al. 2012a), AD prognosis at MCI stage (Coupe et al. 2012b, Eskildsen et al. 2013, Eskildsen et al. 2015, Hett et al. 2018) and at presymptomatic stage (Coupé et al. 2015). First, we will show the performance obtained by HIPP grading compared to HIPP volume – currently considered as the gold standard MRI-based biomarker for AD. Then, we will compare our PBG framework and the proposed extensions with usual biomarkers, but also with recent *state-of-the-art* methods. Finally, we will evaluate the performance of our PBG for predicting AD at presymptomatic stage using a population-based database with a very long follow-up.

2.4.1 Automatic AD Diagnosis

Hippocampal atrophy has long been recognized as an early feature of the degenerative process in AD (Ball et al. 1985). Reductions in hippocampal volume appear to correspond to early memory decline (De Leon et al. 1989). While sensitive, hippocampal degeneration is involved in other dementias, such as vascular dementia (Gainotti et al. 2004), and is known to be part of non-pathological brain aging (Driscoll et al. 2003). Thus, volumetric measurements of the HIPP are limited in their ability to predict the progression of AD (Chupin et al. 2009, Wolz et al. 2011b, Clerx et al. 2013). Evidence suggests that the nature of degeneration in the HIPP and surrounding structures, such as the ENTCOR and parahippocampal gyrus, is different in AD compared to other dementias and different from the changes occurring during normal aging (Devanand et al. 2012). In (Coupe et al. 2012a, Coupe et al. 2012b), we investigated the use of HIPP grading and its surrounding structures in the medial temporal lobe to perform automatic AD diagnosis.

Data used in the main majority of our works dedicated to CAD were obtained from the ADNI database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations as a $60 million, five-year public–private partnership. The primary goal of the ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessen the time and cost of clinical trials.

In this section, the following results were obtained on the 834 baseline scans at 1.5T of the ADNI1 database. The scans were divided into four populations, with an MCI subject considered progressive if he or she converted to AD as of July 2011. This population construction resulted in the four groups composing our dataset: 231 CN, 238 sable MCI (sMCI), 167 progressive MCI (pMCI), and 198 AD. The four constructed groups are the same as those used in (Wolz et al. 2011b). Demographic details of the dataset can be found in Table 8.

Table 8: Demographic details of the ADNI1 dataset used in our experiments.

Group	Population size	% Male	Age ± SD	MMSE ± SD
CN	231	52%	76.0 ± 5.0	29.1 ± 0.9
sMCI	238	67%	74.9 ± 7.7	27.2 ± 2.5
pMCI	167	60%	74.5 ± 7.2	26.4 ± 2.0
AD	198	50%	75.6 ± 7.7	22.8 ± 2.9

Left upper part of Table 9 presents the classification accuracy for AD vs. CN obtained by the proposed imaging biomarkers using LDA as classifier. First, these results show that grading-based biomarkers outperform volume-based biomarkers. Moreover, we can see that ENTCOR-based biomarkers are less efficient than HIPP-based biomarkers. This result was unexpected given that the EC is believed to be affected before the HC in the evolution of the pathology (Frisoni et al. 2010) and thus should be more useful for AD diagnosis. The high intersubject variability related to EC seems to adversely affect the usefulness of this biomarker for the detection of AD. Finally, the combination of HIPP and ENTCOR did not really change results from those obtained with the use of HIPP grade only**.** In order to estimate if the difference between the classification accuracy of biomarkers was significant, we compared the classification results of grading and structure volumes. By using a confidence interval at 95%, all the biomarkers have significantly different accuracy. This result demonstrates the competitive performance of HIPP patch-based grading compared to HIPP volume – considered as the current MRI-based gold standard biomarkers for AD.

Left lower part of Table 9 presents a comparison of our PBG with *state-of-the-art* methods such as HIPP volume, cortical thickness measurements (CTH) and tensor-based methods (TBM) features (Wolz et al. 2011b) evaluayed on the same populations. For AD vs. CN**,** the results obtained with our PBG were similar (91% compared to 89%) to those from the combination of the four methods reported in (Wolz et al. 2011b). Our PBG obtained better results than HIPP volume (Lotjonen et al. 2011), manifold-based learning (Wolz et al. 2011a), CTH (Lerch and Evans 2005), and method based on TBM features (Koikkalainen et al. 2011), although the results from TBM and our PBG were close. The results obtained for HIPP volumes using our PBS (Coupe et al. 2011) and multi-template nonlinear warping (Lotjonen et al. 2011) were also similar (83% compared to 81%). These findings indicate that the both approaches provide similar segmentation accuracies while PBS is much faster especially when using OPAL.

2.4.2 Automatic AD Prognosis at MCI Stage

2.4.2.1 Monomodal Hippocampal Grading

From a clinical perspective, the ability to predict AD (i.e., identifying pMCI vs. sMCI) is more crucial than being able to detect AD (i.e., AD vs. CN). However, prediction is more challenging because the anatomical changes to be identified are subtler at the prodromal phase of the disease and the heterogeneous MCI group includes a mix of individuals, some who will convert to AD and others who will not. As done in previous section, in (Coupe et al. 2012b) we studied the performance of our PBG for AD prediction compared to usual biomarkers on the same populations (Wolz et al. 2011b). The classification results obtained during the comparison are reported in the right part of Table 9.

Table 9: Comparison of classification results on ADNI1 between our PBG and methods studied in (Wolz et al. 2011b). Results shown are obtained using 100 x Leave-N-Out CV. The presented results are the classification accuracy (acc) in %, the sensitivity (sen) in % and the specificity (spe) in %. Best result for each comparison is in bold and underline.

Method	AD vs. CN acc%/sen%/spe%	pMCI vs. sMCI acc%/sen%/spe%
PBG (Coupe et al. 2012b)		
HIPP Volume (based on PBS)	83 / 80 / 85	66 / 65 / 67
ENTCOR Volume (based on PBS)	73 / 71 / 75	63 / 63 / 63
HIPP Grading	90 / 86 / 93	**74 / 73 / 74**
HIPP-ENTCOR Volume	80 / 80 / 81	67 / 66 / 68
HIPP-ENTCOR Grading	**91 / 87 / 94**	73 / 72 / 74
Multi-Method (Wolz et al. 2011b)		
HIPP Volume (based on nonlinear registration)	81 / 81 / 79	65 / 63 / 67
Manifold-based Learning	85 / 87 / 83	65 / 64 / 66
Cortical Thickness	81 / 89 / 71	56 / 63 / 45
Tensor-based method	87 / 90 / 84	64 / 65 / 62
All combined	**89 / 93 / 85**	**68 / 67 / 69**

This comparison shows that results obtained by our PBG were clearly better than those from all the methods compared in (Wolz et al. 2011b) as well as their combination (74% compared to 68%). This outcome highlights the potential of our PBG for AD prediction by enabling the detection of subtle anatomical changes caused by AD at the early stages of the pathology. Although the prediction rate obtained is not yet suitable for clinical use, the recent progress of MRI-based biomarkers on this challenging classification problem is encouraging. Before PBG publication, the highest success rate was only around 56% on the ADNI database using advanced VBM-like analysis (Davatzikos et al. 2011). It is also encouraging to note that the improvements brought by our PBG were not obtained at the expense of computational complexity since PBG requires only linear registration and its computational time is few seconds per subject using OPAL implementation.

2.4.2.2 Monomodal Grading Extensions

Since its introduction, PBG has been intensively studied and several improvements have been proposed. First, in (Tong et al. 2014), the patch comparison was achieved using multiple instance learning (MIL). Second, in (Liu et al. 2012, Tong et al. 2017a), the weights were estimated using a sparse-based minimization and the grading is extended to the entire brain. Finally, as already mentioned, we proposed multifeature grading using texture maps (Hett et al. 2017) and whole GM grading using ensemble grading (Komlagan et al. 2014). Table 10 presents the comparison of all these patch-based grading strategies. Moreover, in this manuscript I decided to add a recent DL method (Suk et al. 2017). First, we compared our different HIPP grading methods (Coupe et al. 2012b, Hett et al. 2017) with other grading methods focused on HIPP (see the upper part of Table 10). This comparison shows that our multifeature grading provides the best results among HIPP grading methods and that the original HIPP grading is still competitive for sMCI vs. pMCI several years after its publication compared to MIL (Tong et al. 2014) grading or recent sparse-based grading (Tong et al. 2017a). These results also demonstrate that texture maps provide valuable information especially for AD vs. CN.

Table 10: Comparison with state-of-the-art methods on the ADNI1 dataset. All the results are expressed in accuracy. The underlined names are the methods that we proposed. The best results are in bold and underlined.

Method	Registration	Features	AD vs. CN ACC in %	sMCI vs. pMCI ACC in %
Hippocampus				
Original Grading (Coupe et al. 2012b)	Affine	Intensity	88.0	71.0
MIL Grading (Tong et al. 2014)	Affine	Intensity	89.0	70.0
Sparse-based Grading (Tong et al. 2017a)	Affine	Intensity	-	66.0
Sparse-based Grading (Tong et al. 2017a)	Non-linear	Intensity	-	69.0
Multifeature Grading (Hett et al. 2017)	Affine	Texture	**91.3**	71.1
Whole brain				
Ensemble Grading (Komlagan et al. 2014)	Nonlinear	GM map	-	**75.6**
Sparse-based Grading (Tong et al. 2017a)	Affine	Intensity	-	67.0
Sparse-based Grading (Tong et al. 2017a)	Non-linear	Intensity	-	75.0
Sparse Ensemble Grading (Liu et al. 2012)	Non-linear	GM map	90.8	-
Deep Ensemble Learning (Suk et al. 2017)	Non-linear	GM map	91.0	74.8

At the lower part of Table 10, we compare the performance of our GM ensemble grading method based on the whole GM (Komlagan et al. 2014) with methods using also the whole brain. Results show that our whole GM grading provides the best result for sMCI vs. pMCI, better results than a last advanced DL method (Suk et al. 2017). In addition, for AD vs. CN, we can note that our multifeature HIPP grading method based on a simple affine registration obtained similar results than more complicated methods using whole brain and requiring nonlinear registration (Liu et al. 2012, Suk et al. 2017). Finally, for sMCI vs. pMCI, the use of the whole brain provides better results indicating that relevant information is present outside the HIPP.

2.4.2.3 Multimodal Hippocampal Grading

As previously mentioned, beside anatomical MRI, the use of diffusion MRI has been proposed to detect the first signs of microstructure alterations caused by AD. Several studies used DTI to detect modifications of diffusion parameters into the whole white matter (O'Dwyer et al. 2012, Dyrba et al. 2015). Others studies showed modifications of diffusion parameters into specific structures such as corpus callosum (Nir et al. 2013, Wang et al. 2015), fornix (Liu et al. 2011), cingulum (Nir et al. 2013) and hippocampus (Rose et al. 2008). More advanced diffusion MRI studies using have been proposed to extract features describing axonal fibers alterations (Liu et al. 2011, Wee et al. 2012, Prasad et al. 2015). Finally, it has been shown that HIPP mean diffusivity (MD) is correlated to pathology progression and thus could be used as an efficient biomarker of AD (Müller et al. 2005, Fellgiebel et al. 2006, Fellgiebel and Yakushev 2011). Consequently, in (Hett et al. 2018), we proposed to study HIPP alterations occurring at different stages of impairment severity using structural MRI and diffusion MRI modalities.

Data used in (Hett et al. 2018) were obtained from the ADNI2 dataset that provides anatomical MRI and diffusion MRI scans (see Table 11). This dataset includes AD, CN and MCI divided into 2 stages: early MCI (eMCI) and late MCI (lMCI). To evaluate the performance our multimodal patch-based grading, we first compared it with several biomarkers over the whole HIPP.

Table 11: Description of the multimodal (structural MRI + diffusion MRI) ADNI2 dataset used in (Hett et al. 2018)

Group	Population size	Gender (F/M)	Age ± SD
CN	52	29/23	72.6 ± 5.9
eMCI	65	39/26	73.0 ± 7.7
lMCI	34	21/13	73.5 ± 6.6
AD	38	20/18	73.8 ± 8.7

Results of this comparison are represented in Table 12. First, for CN vs. AD, our PBG on T1w was the modality given the best results with an AUC of 93.1%. The performance of PBG based on T1w could be explained by the higher resolution of T1w compared to DTI allowing to better capture structure alterations. On the other hand, these results show that PBG based on MD was the best for the eMCI vs. lMCI with an AUC of 66.7%. Therefore, at early stages of AD, microstructural information provided by DTI seems to be useful. Finally, our novel multimodal PBG provided the best results for eMCI vs. lMCI with an AUC of 66.7% and the second best result for AD vs. CN with an AUC 92.2%. This demonstrates the interest of using MPBG in distinguishing the different MCI stages.

2.4.2.4 Multimodal Hippocampal Subfield Grading

Although majority of proposed biomarkers to detect AD are based on the whole HIPP, this structure is complex and not homogeneous. Different HIPP subfields segmentation protocols have been proposed (Yushkevich et al. 2015a). However, as already mentioned in Chapter 2, HIPP can mainly be divided into the subiculum, the cornu ammonis (CA1/2/3/4), and the dentrate gyrus (DG) (Winterburn et al. 2013). The CA1 subfield is the biggest area in the hippocampus. It is composed by different layers called the stratum radiatum (SR), the stratum lacunosum (SL), the stratum molecular (SM) and the stratum pyramidale (SP). Several MRI studies demonstrated that subfields are impacted differently according to AD stages. In (La Joie et al. 2013), the authors showed that the CA1 is the most impacted subfield in advanced AD. In (Apostolova et al. 2006), the authors showed that CA1 and subiculum are more impacted than the others subfields in late MCI and advanced AD stages. A study based on ultra-high resolution MRI at 7T showed that CA1SR-L-M atrophy appears when CA1SP or global HIPP atrophy are not detectable yet that demonstrates the specific atrophy of CA1SR-L-M in first stage of AD (Kerchner et al. 2012). Moreover, a study based on animal model showed that the earliest affected hippocampal region is the subiculum (Trujillo-Estrada et al. 2014). Therefore, in (Hett et al. 2018), we proposed to perform AD prognosis based on MPBG of hippocampal subfields. The hippocampal subfields were segmented with our PBS method called HIPS described in the previous chapter (Romero et al. 2017b).

Table 12: Comparison of the considered hippocampal biomarkers on the multimodal ADNI2 dataset. Results shown are obtained using Leave-One-Out CV and a LDA as classifier. The presented results are the Area Under the curve (AUC) in %. Best result for each comparison is in bold and underline.

Method	AD vs. CN AUC%	eMCI vs. lMCI AUC%
HIPP Volume	86.7	55.2
HIPP MD	83.0	49.2
HIPP T1w Grading (Coupe et al. 2012b)	**93.1**	64.9
HIPP MD Grading (Hett et al. 2016)	89.5	**66.7**
HIPP MPBG (Hett et al. 2018)	92.2	**66.7**

Figure 21 shows the classification results for CN vs. AD (A) and eMCI vs. lMCI (B). For CN vs. AD, the subfield providing the most discriminant volume was the CA1SR-L-M with an AUC of 86.0%. Moreover, the subfield providing the most discriminant mean MD was the subiculum with an AUC of 88.1%. Mean MD of subiculum was the only non-grading biomarker performing better results than whole hippocampus volume – considered as the current gold standard. The subfields providing the most discriminant MPBG were the CA1SP, CA1SR-L-M and the subiculum with an AUC of 91.7%. All the MPBG subfields biomarkers performed better than global HIPP volume (see horizontal red dotted line in Figure 21.A). However, for CN vs. AD, none of the MPBG subfield biomarkers outperformed MPBG estimated over the whole HIPP (horizontal bleu dotted line) that obtained an AUC of 92.2% (see Table 12). It seems that for advanced AD stages, when brain alterations are important, finer analysis at subfield scale does not bring any additional information.

On the other hand, for eMCI vs. lMCI comparison, the subiculum provides the best results for all considered biomarkers (i.e., volume, mean MD ad MPBG), better than global volume or global MPBG estimated over the whole HIPP (HIPP MPBG obtained 66.7%, see Table 12). As shown in Figure 21.B, subiculum obtains an AUC of 66.0% for the volume, 63.0% for mean MD, and 72.0% for MPBG. Interestingly, all three biomarkers demonstrate the efficiency of the subiculum in distinguishing between early and late stages of MCI.

To sum up, our volumetric study showed that CA1SR-L-M presents the strongest atrophy at the advanced stage of AD. However, for the early stages of the pathology, our comparison showed that the subiculum is clearly the most discriminant structure. It is important to note that these results are line with previous studies that analyzed hippocampal subfield alteration caused by AD.

First, at advanced stages of AD, several studies showed that the CA1 and the subiculum were the two subfields impacted by the strongest atrophy (Apostolova et al. 2006, La Joie et al. 2013, Carlesimo et al. 2015). Furthermore, studies using ultra-high field at 7T indicated that CA1SR-L-M is the subfield impacted by the most significant atrophy at an advanced stage of the pathology (Kerchner et al. 2010, Kerchner et al. 2012).

Second, at early stages of AD, studies based on animal model demonstrated that subiculum is the earliest hippocampal region affected by AD (Trujillo-Estrada et al. 2014). Moreover, as already mentioned, *post-mortem* studies showed that the hippocampal degeneration in early stages of AD is not uniform. After ENTCOR alterations, the pathology spreads to the subiculum, CA1SR-L-M, CA2-3 and finally the CA4 and DG subfields (Braak and Braak 1997, Braak et al. 2006). These observations are consistent with our results obtained using two different MRI modalities and three different types of biomarker.

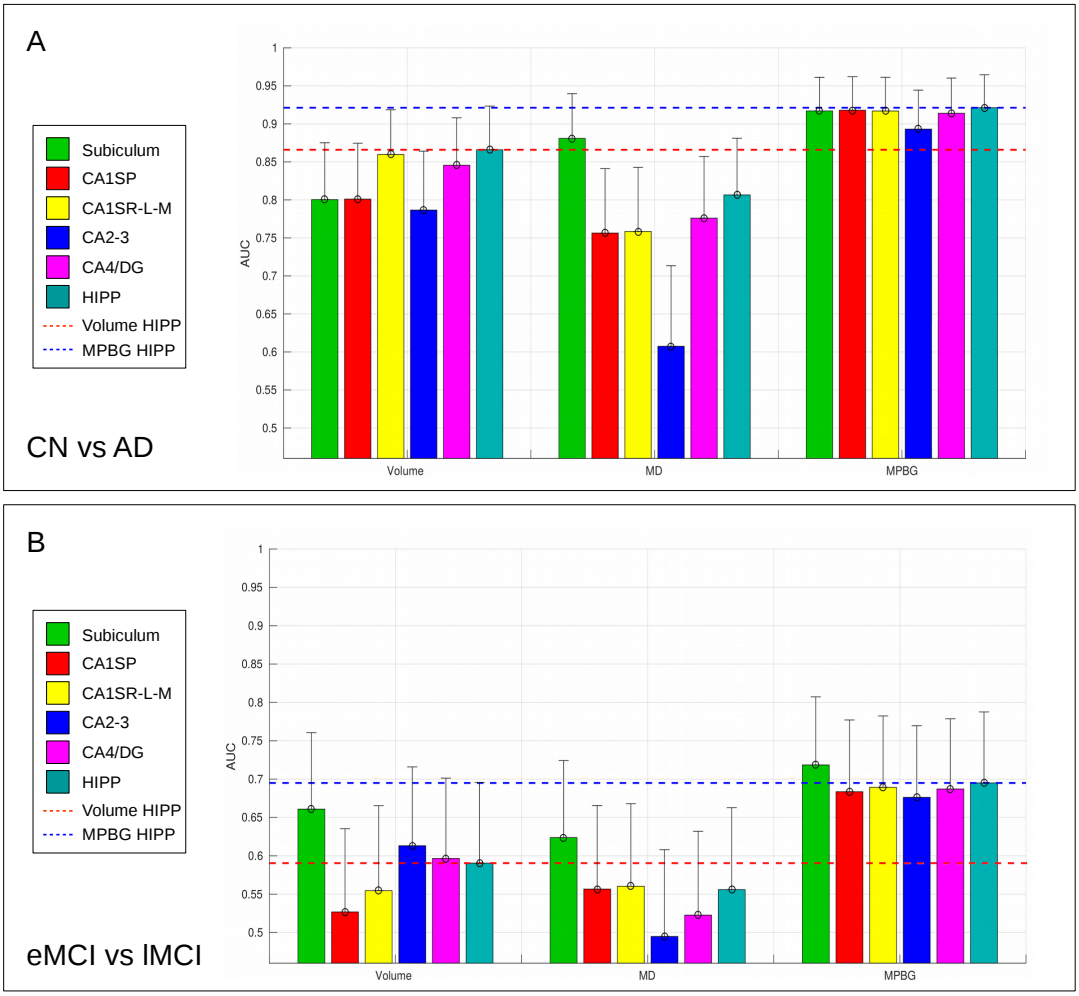

Figure 21: Comparison of multimodal biomarkers over hippocampal subfields in term of AUC. Results of subfields are grouped by biomarker (i.e., volume, mean MD and MPBG). Moreover, two comparisons are proposed, A: CN versus AD and B: eMCI versus lMCI. Finally, horizontal dotted lines represent results obtained over the entire HIPP for volume in red and MPBG in bleu.

2.4.3 AD prognosis at presymptomatic stage

While structural imaging markers based on MRI are considered sensitive enough to detect very early stage of disease (Frisoni et al. 2010, Cuingnet et al. 2011, Braskie and Thompson 2014), the current model assumes that their modification occurs after the apparition of the first symptoms (Frisoni et al. 2010, Jack et al. 2013) (see Figure 15). This assumption mainly comes from observations based on MRI-based studies were used datasets had a relatively short follow-up period before diagnosis (around 3-5 years). However, it is now well admitted that pathological changes to the brain occur decades before the first symptoms appear in AD. Thus, efforts to identify subjects in the prodromal phase of AD have recently shifted to the presymptomatic phase of the disease. In this context, it is highly desirable to use population-based cohorts that include healthy elderly with longer follow-up. In addition, population-based cohorts give us the opportunity to avoid the potentially biased selection associated with recruitment in memory clinics and enable the study of imaging and neuropsychological parameters at the presymptomatic stage (silent phase) of the disease.

In (Coupé et al. 2015), we proposed to evaluate MRI-based imaging biomarkers on a dataset from a population-based cohorts of healthy elderly subjects with long follow-up (12 years) to study imaging and neuropsychological parameters over the course of the presymptomatic phase of AD. During the 12 years follow-up period, some subjects will convert to AD after enrollment into the study (7 years on average). This long follow-up enables to track evidence of neurodegeneration at least 7 years before clinical diagnosis using MRI.

In this work, we used subjects from the Bordeaux site of the Three-City (Bdx-3C) dataset, a prospective population-based cohort designed to assess the risk of dementia and cognitive decline due to vascular risk factors (3C-Study-Group 2003). In this longitudinal dataset, neuropsychological tests were performed by trained psychologists at baseline and at 2, 4, 7, 10 and 12 years. MRI scans were acquired at baseline and at 4 and 9 years. The neuropsychological evaluation consisted of several tests performed at each follow-up, always including the MMSE (Folstein et al. 1975) to evaluate global cognitive function. During follow-up, subjects suspected of dementia were seen at home by a neurologist who established a preliminary diagnosis. After that, definitive diagnosis was made by a panel of independent neurologists to obtain a consensus on the diagnosis (McKhann et al. 1984)**.**

In the Bdx-3C study, 663 subjects have a baseline MRI scan. By using baseline scans of all the CN subjects who remain free of dementia during 12 years, we built a stable CN (sCN) group while subjects who converted to AD during the last 3 follow-ups were used to build the converter CN (cCN) group. We do not use CN subjects who converted at 2-year and 4-year follow-up in order to build a homogeneous group composed of subjects far from conversion. Therefore, Bdx-3C dataset was divided into two groups of 309 sCN and 37 cCN. Demographic details of both Bdx-3C groups are provided in Table 13. We found statistical difference between sCN and cCN groups for age (p=0.0001) but not for gender (p=0.17). At baseline, no statistical difference was observed between groups for a global cognitive scale (i.e., MMSE) using generalized linear model. This indicates that cCN subjects are in a silent phase.

We estimated the mean time before AD diagnosis for the cCN group at 103 months (i.e., 8.6y) using the baseline MRI and the follow-up exams where AD diagnosis was established. However, low frequency of follow-up exams might introduce a bias by overestimating this duration. Therefore, we also computed an estimated average time before conversion to compensate for this bias. To do that, we used the average time between the closest follow-up exam before AD diagnosis and the follow-up exam when AD diagnosis was established. Using this procedure, the mean time before the estimated conversion to AD was 84 months (i.e., 7y). Finally, we used AD and CN groups of ADNI1 dataset as training library (see Table 8) to perform HIPP grading of the sCN and cCN groups from Bdx-3C data.

Table 13: Demographic details of the stable CN and converter CN of the Bdx-3C dataset used as testing images.

Group	Population size	% Male	Age (SD)	MMSE (SD)
Stable CN (Bdx-3C)	309	41%	72.7 (3.9)	28.4 (1.2)
Converter CN (Bdx-3C)	37	30%	75.4 (3.9)	27.9 (1.4)

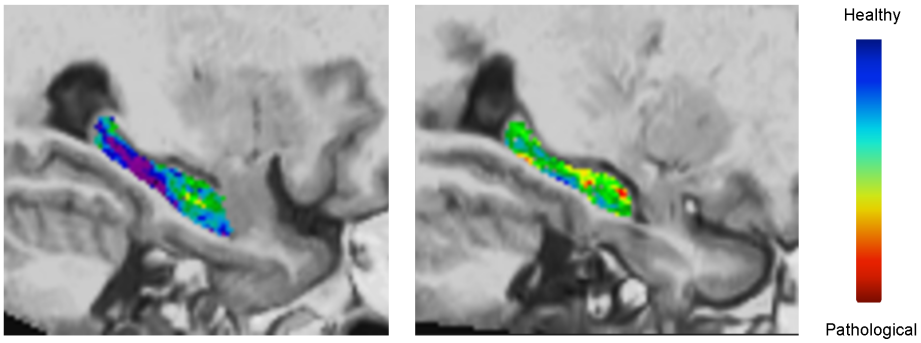

Figure 22: Examples of hippocampal grading maps obtained on the Bdx-3C dataset for a sCN subject and a cCN subject 7.5 years before conversion to AD. The blue-purple hippocampal grading map of the sCN subject indicates that a majority of CN-like anatomical patterns in this subject has been detected. The green areas in the hippocampal grading map of the cCN subject indicate the hippocampal structural patterns of this subject are equally similar to the anatomy of AD and CN of the ADNI1 training subjects. Finally, red regions indicate that AD-like anatomical patterns have been found in this cCN subject 7.5 years before conversion to AD.

First, we proposed visual assessment. Figure 22 shows typical examples of hippocampal grading maps for age-matched sCN and cCN subjects. The hippocampal grading map of the sCN subject (mainly blue-purple color) indicates that the majority of the anatomical patterns of the subject are highly similar to anatomical patterns found in the ADNI1 CN training subjects. For the cCN subject, a number of red focal regions appear, indicating that these hippocampal regions exhibit high structural similarity with the pathological pattern of ADNI1 AD training subjects. This indicates that in these areas, AD-like anatomical patterns were found in this cCN subject and thus that signs of neurodegeneration were detected 7.5 years before conversion to AD. Second, to study the AD prognostic detection accuracy at presymptomatic stage (i.e. 7y before conversion or 8.6y before diagnosis) of several biomarkers. The considered biomarkers were HIPP grade, HIPP volume, MMSE and patch-based MMSE estimation. Recently, several studies investigated age prediction (Franke et al. 2012) and clinical scores prediction (Stonnington et al. 2010, Zhang et al. 2012) using MRI content. The interest of such image-based approaches is growing for early pathology detection and longitudinal follow-up. As previously explained, the proposed PBG framework can be adapted to other problems than clinical status estimation. Instead of using clinical status of ADNI1 training subjects as input to estimate clinical status of Bdx-3C subjects, MMSE scores of ADNI1 training subjects can be used to perform a patch-based MMSE estimation.

Table 14 shows the classification performance obtained with a LOOCV using a LDA as classifier. HIPP grading obtained the highest classification accuracy (72.5%), followed by patch-based MMSE estimation (68.8%), HIPP volume (58.1%) and the MMSE (56.9%). Therefore, HIPP grading provided a gain of 14.4 percentage point (pp) in term of prognosis accuracy. In addition, HIPP grade obtained the best sensitivity and specificity with a gain of 10.9 pp for sensitivity and of 14.9 pp for specificity. According to the McNemar's test, only classification based on imaging biomarkers performed significantly ($p \leq 0.05$) better than a random classifier. The HIPP volume was marginally significant ($p=0.04$) while HIPP grading and patch-based MMSE estimation were highly significant ($p<0.0001$). The patch-based MMSE clearly improves prognosis accuracy compared to clinical MMSE scores with a gain of 11.9 pp in terms of ACC and 21.2 pp in terms of AUC. The patch-based MMSE AUC is similar to values obtained with hippocampal grading while ACC is slightly lower.

*Table 14: Classification results based on LDA classifier with values adjusted for age difference. A leave-one-out cross-validation procedure was used. Results are reported for ACC=Accuracy, SEN=Sensitivity, SPE=Specificity, p-value of the McNemar's test, AUC=Area under the ROC curve and CI=95% Confidence Interval. * Classification performance estimated as significantly better than random classifier using the McNemar's test. The best results are in bold font. These results are obtained on cCN at 7y before conversion to AD and at 8.6y before AD diagnosis.*

	cCN (37) vs. sCN (309)				
	ACC %	SEN %	SPE %	McNemar's test	AUC [95% CI]
MMSE scores	56.9	46.0	58.3	0.08	52.0 [40.4-61.6]
Hippocampal volume	58.1	54.0	58.6	0.04*	64.6 [54.4-72.3]
Patch-based MMSE estimation	68.8	59.5	69.9	**p<0.0001***	**73.2 [63.8-81.5]**
Hippocampal grading	**72.5**	**64.9**	**73.5**	**p<0.0001***	73.0 [63.6-81.3]

2.5 Discussion

In this second chapter, we presented our PBG framework and the proposed extensions. Based on our PBG method, we developed computer-aided diagnosis tools that we applied to AD diagnosis and prognosis. During our experiments, we showed that the proposed PBG framework obtained high prognosis accuracy even at presymptomatic phase. Recent, independent studies on the ADNI1 dataset confirmed our results and showed that PBG provides higher AD prognosis accuracy than *state-of-the-art* methods (Tong et al. 2014, Tong et al. 2017a).

In (Coupe et al. 2012a, Coupe et al. 2012b), we showed that PBG grading biomarkers provides competitive results for early detection of AD on the ADNI1 dataset compared to conventional methods such as HIPP volume, CTH, and method based on Tensor-Based Morphometry features. The comparison of PBG biomarkers in the context of early detection demonstrated the high potential of the proposed framework for this key clinical problem. Although the obtained prediction rate of 74% is not yet suitable for clinical use, the recent progress of MRI-based biomarkers on this challenging classification problem is encouraging. Moreover, the simplicity of the PBG framework results in a robust pipeline; the processing failure rate was less than 1.7% at the linear registration step — a much lower failure rate in great contrast to the 13% obtained for the CTH method presented in (Wolz et al. 2011b). Finally, in (Eskildsen et al. 2015) we showed the advantage of combining CTH and grading features for AD prediction. We demonstrated the complementary of cortical and hippocampal biomarkers to improve classification accuracy at MCI stage.

In (Komlagan et al. 2014), we proposed to extent HIPP grading to the whole gray matter using ensemble-based framework and an automatic selection of relevant anatomical areas. Compared to HIPP grading, we improved the accuracy of our PBG using this strategy (see Table 10). It is interesting to note that directly using all the anatomical sub-ensembles (i.e., without SLR-based anatomical areas selection) provided worst results than using only HIPP grading. However, when selecting the most relevant anatomical sub-ensembles an important increase is observed. This indicates that areas other than hippocampus seem to be impacted at MCI stage. Thus, automatic *a posteriori* selection of these areas instead of using predefined ROIs leads to higher accuracy. As shown in Table 10, our ensemble-based grading achieved better prognosis accuracy than other *state-of-the-art* methods validated on the same ADNI1 database. So far, our ensemble-based grading method remains very competitive compared to recently proposed methods based on deep learning (Suk et al. 2017).

In (Hett et al. 2017), we investigated the potential of using texture information based on Gabor filters to improve patch- based grading method performance. We compared our new texture-based grading biomarker with *state-of-the-art* approaches and we showed the high potential of the proposed method for AD diagnosis (see Table 10). However, according to our comparison and as we showed it in (Komlagan et al. 2014), whole brain methods enable a better classification of sMCI vs. pMCI. Hence, in further works, we will investigate the extension of our texture-based grading framework to whole brain analysis.

In (Hett et al. 2016, Hett et al. 2018), we studied the capabilities of MPBG by combining MRI and DTI-based grading. We showed that MD grading can be useful at the earliest stages of AD and that multimodal grading yielded to the most stable results and thus it can a good biomarker candidate (see Table 12). Moreover, in (Hett et al. 2018), we investigated the use of hippocampal subfields MPBG to better distinguish AD stages. In this study, the comparison based on MD, volume and MPBG biomarkers showed that the subiculum is the most discriminant structure in the earliest stage of AD. Moreover, we showed that the CA1SR-L-M presented the strongest atrophy in the advanced stage of AD. These results are in accordance with previous literature (Apostolova et al. 2006, Kerchner et al. 2010, Kerchner et al. 2012, La Joie et al. 2013, Trujillo-Estrada et al. 2014).

In (Coupé et al. 2015), we presented a study to assess the predictive value of HIPP grading on incident dementia in cognitively intact patients more than 7 years before conversion. We showed that using HIPP grading for early prognosis exhibited higher classification accuracy than HIPP volume with a gain of 14pp. This result on the Bdx-3C population is in line with our result obtained on the ADNI1 dataset with subjects who converted to AD in 18 months on average. In (Coupe et al. 2012b), we obtained 74% of ACC using HIPP grade and 66% of ACC using HIPP volume on the ADNI1 dataset. These results support the hypothesis that, although hippocampal volume has been found to be an efficient imaging biomarker on subjects close to AD diagnosis (Wolz et al. 2011b, Ewers et al. 2012), hippocampal volume loss is not sufficiently sensitive to enable automatic classification at a very early preclinical stage. The use of more sophisticated measures, such as our PBG, appears to be crucial for the identification of AD at presymptomatic phase. From a pathophysiological point of view, the performance of the HIPP grading supports the hypothesis of HIPP anatomical alterations 7 years before conversion to AD. Taken together these results are consistent with the idea that clinical AD is preceded by a long asymptomatic phase, which is characterized by progressive functional, metabolic and anatomical brain alterations.

In addition, in (Coupé et al. 2015), we performed patch-based MMSE estimation using our HIPP grading framework. Recently, a whole brain extension of this strategy has been successfully used for age estimation and clinical scores estimation (e.g., ADAS) (Wu et al. 2016a). This independent study shows that our patch-based framework can be generalized to any attributes of the training subjects. Finally, we recently proposed on longitudinal version of our PBG in (Sanroma et al. 2017). In this work, the grading is estimated between two time points using the longitudinal subject self-similarity.

So far, we used PBG grading mainly for AD diagnosis and prognosis. However, a recent independent study applied whole brain grading to differential diagnosis between AD, vascular dementia (VaD), dementia with Lewy bodies (DLB) and frontotemporal dementia (FTD) (Koikkalainen et al. 2016). The authors showed that, compared to volume, TBM, VBM and Manifold learning features (Wolz et al. 2011b), PBG was the most competitive biomarker to detect AD and FTD. More importantly, PBG was the best feature to differentiate AD vs. FTD, AD vs. DLB and FTD vs. DLB. These results indicate that PBG can help to distinguish different types of dementia which is a crucial clinical question. These results have been recently confirmed in (Tong et al. 2017b), where PBG demonstrated very competitive results for 5-class differential diagnosis of AD, FTD, DLB, VaD and SMC (i.e., subjects with memory complaints who did not meet the criteria for MCI). For our part, we plan to apply PBG to multiple sclerosis to predict pathology evolution at the earliest stages of the pathology.

Chapter 3 Towards Brain Monitoring

Abstract: In this last chapter, we will present the tools developed to perform brain monitoring. Such monitoring is possible only when two elements are available – a tool to measure the studied parameter and a standard model to know when the obtained measurement is normal or not. First, we will describe the volBrain pipeline, the tool that we proposed to perform automatic quantitative brain analysis. This pipeline provides the volume of anatomical brain structures in a fast and robust manner and thus it enables to process very large databases. Second, we will present the standard models for anatomical brain structures that we estimated to determine when structure volumes are normal or not. Thanks to the new paradigm of Big Data sharing in neuroimaging, we developed standard models across the entire lifespan based on a massive number of freely available MRI. Finally, we will present the proposed open access web-platform that integrates the volBrain pipeline and our lifespan models. This service is accessible through our volBrain web-platform as a software as a service. To conclude, we will discuss about the platform usage after 3 years of experience.

3.1 Introduction

Every day in hospitals over the world, we measure physiological parameters to monitor body development, to detect abnormalities or to establish a diagnostic. This monitoring is possible only when two elements are combined. First, we need a tool to measure the studied parameters (*e.g.,* blood pressure or heart rate). This tool has to be convenient, standardized and easily accessible. Second, we need a standard model to know when the obtained measurement is normal or not compared to reference cases. While for many physiological parameters such normative values exist, they do not exist for volume of brain structures (see Figure 23). Recently, to overcome these two limitations, we proposed to develop measurement tools and their corresponding standard models.

First, we needed efficient tools to perform quantitative brain analysis enable to process the massive number MRI produced every day. To move from qualitative to quantitative analysis, we had to develop fast, robust, accurate and scalable methods to automatically extract useful information from medical images. The main challenge was to propose accurate pipelines enable to process a large number of MRI in a reasonable time and with a low failure rate**.** Therefore, we developed new fast and robust pipelines for brain segmentation based on our PBS framework (Manjon and Coupe 2016).

Second, to determine when measurements are normal or not compared to reference cases, we needed standard models. In the literature, brain development has been studied previously but only over restricted periods (e.g., childhood or old age) and using a restricted number of scans (usually several hundreds). Thanks to the new paradigm of Big Data sharing in neuroimaging, we have been able to develop standard models for brain monitoring using a massive number of freely available MRI. Therefore, we proposed standard models to study human brain trajectory from few months of life to advanced ages (Coupe et al. 2017). We proposed also an AD model based on the same strategy (Coupé et al. 2018).

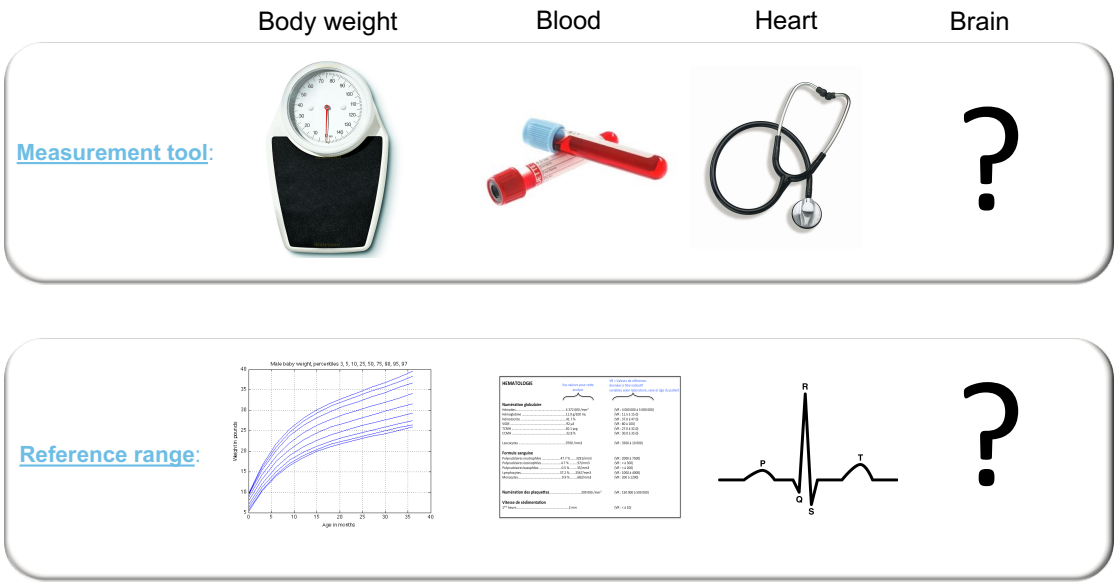

Figure 23 : Example of usual physiological monitoring. While for many physiological parameters measurement tools and normative values exist, they do not exist for volume of brain structures.

Finally, we integrated our tools and models into a web platform in full open access. Most of developed pipelines for MR image analysis are software packages that need to be installed and configured. Installation and configuration steps can be complicated and require experimented people and computational resources. To limit the need for infrastructure, installation of software on grid computing has gained in popularity. Unfortunately, these platforms are not all open access and are difficult to use for non-experts. Moreover, most of the available platforms offer almost the same tools (e.g., FSL or/and Freesurfer) what yields to redundant services**.** Therefore, we proposed our original tools and the corresponding standard models in full open access as Software as a Service (SaaS) through the volBrain platform.

In this chapter, we will present these three contributions and the obtained results. First, we will detail the proposed pipeline to perform brain segmentation in an accurate and robust manner compared to *state-of-the-art* methods. Second, we will present how we estimated standard values across the entire lifespan. Moreover, we will detail the pathological models based on AD population that we used to estimate the point of divergence between normal and pathological brains across the entire lifespan. Finally, we will describe the volBrain web-platform and the encouraging results obtained by this system after 3 years of experiences. The several integrated pipelines will be briefly detailed.

3.2 Methods

Automated and reliable quantitative MRI-based brain image analysis has a huge potential to objectively help in the diagnosis and follow-up of many neurological diseases. Specifically, MRI brain structure volumetry is being increasingly used to understand the nature and evolution of those diseases. Over the past decades, manual segmentation has been the method of choice to accurately analyze specific brain structures. However, this task is tedious and time consuming, limiting its use in clinical practice. To make easier this quantification process, many automatic tools have been proposed. Consequently, the brain segmentation problem one of the most intensively studied topics over the past years. Moreover, the recent massive increase of neuroimaging data to process have motivated the development of innovative approaches able to address challenges related to this new "Big Data" paradigm (Van Horn and Toga 2014). Therefore, efficient, automatic, robust and reliable methods for automatic brain analysis will play a major role in the near future.

Several software tools have been developed to automatically estimate brain structure volumes using different strategies. For example, the SPM software (Ashburner 2012) is a widely used tool to analyze global GM or WM alterations. Voxel-Based Morphometry (VBM) toolbox (an extension of SPM) has also been used to measure local GM atrophy (Ashburner and Friston 2000). To perform more specific volume measurements over anatomical structures, tools like the FSL package (Jenkinson et al. 2012) or Freesurfer (Fischl 2012) have been developed. FSL is a comprehensive library of analysis tools for functional MRI, anatomical MRI and DTI brain imaging data. One of these tools, called FIRST (Patenaude et al. 2011), is able to automatically segment subcortical brain structures. Similarly, the Freesurfer pipeline can be used for volumetric segmentation, cortical surface reconstruction and cortical parcellation. Freesurfer has been used in numerous studies despite its high computational burden.

3.2.1 The volBrain Pipeline

In (Manjon and Coupe 2016), we proposed a new software pipeline for volumetric brain analysis. This pipeline provides automatically volumetric brain information of different anatomical structures. In the following, the different parts of the volBrain pipeline will be described and a comparison with *state-of-the-art* methods will be presented. The volBrain pipeline proposes segmentations at different scales:

- Intracranial cavity extraction
- WM, GM and Cerebrospinal Fluid (CSF) tissue classification.
- Cerebrum, cerebellum and brainstem segmentation (separating left from right cerebrum and cerebellum).
- Lateral ventricles and subcortical GM structures segmentation (putamen, caudate, pallidum, thalamus, hippocampus, amygdala and accumbens).

After several preprocessing steps (denoising, inhomogeneity correction and registration into the MNI space), all the segmentations with the exception of tissue volumes are based on different adaptations of our PBS framework (see the pipeline steps on Figure 24).

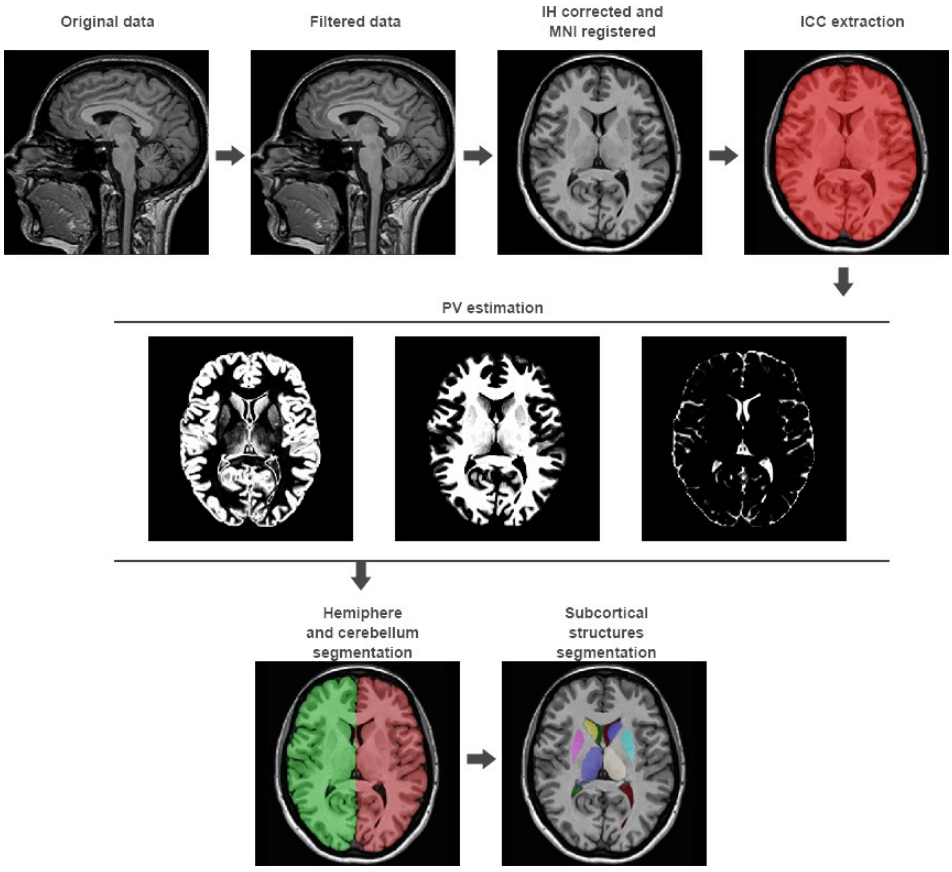

Figure 24: The volBrain processing pipeline. In the first row, the preprocessing is presented. It consists in a non-local noise reduction filter (Manjon et al. 2010c), inhomogeneity correction (Tustison et al. 2010), MNI space affine registration (Avants et al. 2011), intensity normalization and ICC extraction (Manjon et al. 2014). In the second row, the result of the global tissue estimation (GM, WM and CSF) is shown. In the third row, the result of the macrostructures and subcortical structures segmentation is presented.

3.2.1.1 Intracranial Cavity Extraction

After preprocessing, the first step of our pipeline is the intracranial cavity (ICC) extraction. As previously mentioned in Chapter 1 during the BEaST description, PBS can be used for brain segmentation. In the volBrain pipeline, we integrated our intracranial cavity extraction called NICE (Manjon et al. 2014) that is an evolution of the BEaST technique enabling faster and more accurate results and that segment ICC instead of brain. In NICE, we extended BEaST definition by including all external CSF (thus covering total CSF of ICC) and therefore selecting most of the intracranial cavity volume. We have not included other intracranial tissues in our mask definition such as dura, exterior blood vessels or veins because they are normally of no interest for brain analysis. This mask definition has been used to estimate the Total Intracranial Volume (TIV) in many methods such as SPM8 or VBM8 methods in order to normalize brain tissue volumes. In (Manjon et al. 2014), we showed that NICE provides significantly better results than VBM8 and BEaST. Moreover, we demonstrated that NICE has a better reproducibility than these two methods. Finally, NICE performs intracranial cavity extraction in less than 5 minutes.

3.2.1.2 Tissue Classification

Once ICC extraction is done, volBrain pipeline classifies the main intracranial tissues (i.e., WM, GM and CSF). Classical approaches to estimate tissue classification are based on mixture of Gaussians algorithm (Wells et al. 1996), fuzzy C-means clustering (Ahmed et al. 2002) or Markov Random Field (MRF) models (Scherrer et al. 2009). Moreover, a common feature of those methods is the use of *a priori* information in the form of spatial probability maps (e.g. SPM). All these methods assign at each voxel a membership degree or probability to belong to specific tissue rather than calculate the actual amount of each tissue within each voxel. For this reason, some authors used the concept of partial volume coefficients (PVC) to represent the actual amount of every tissue within each voxel. Therefore, we used PVC-based tissue classification in the volBrain pipeline (Manjón et al. 2010). This method combines a MRF model and non-local means filtering to reduce random noise in PVC estimation.

3.2.1.3 Hemisphere Segmentation

Although, the global amount of WM, GM and CSF within the ICC may be an interesting biomarker for quantitative brain analysis, diseases may present local alterations instead of global ones. In addition, the assessment of brain structure asymmetries may be also interesting to study normal/abnormal brain development and to detect alterations due to some neurological diseases. Therefore, segmentation of structures such us cerebrum, cerebellum, brainstem and brain hemispheres might be important to assess brain asymmetry. Several automatic strategies have been developed for hemisphere segmentation (Prima et al. 2002, Mangin et al. 2004, Zhao et al. 2010). Recently, we presented a novel and competitive approach called NABS (Romero et al. 2015) that is based on our PBS method. This method demonstrated competitive performance compared to *state-of-the-art* methods. Moreover, NABS provides left/right hemispheres, left/right cerebellum and brainstem segmentation in less than 1 minute. Therefore, we used NABS in our volBrain pipeline.

3.2.1.4 Structure Segmentation

Finally, it may be also interesting to measure local volumes at a finer scale since many pathologies affect specific areas of the brain. For instance, the hippocampal volume and the lateral ventricles volume have been shown to be early biomarkers of Alzheimer disease as discussed in Chapter 2. To segment subcortical nuclei, several automatic methods have been proposed using deformable models or atlas/template-warping techniques (Collins et al. 1995). As explained in Chapter 1, multi-atlas label fusion segmentation techniques have gained popularity recently because they can combine multiple atlas information to minimize mislabeling from inaccurate affine or non-linear registration (Heckemann et al. 2006, Collins and Pruessner 2010, Lotjonen et al. 2010). However, we showed that our PBS framework provides *state-of-the-art* results in a drastically reduced computational. Therefore, the volBrain pipeline integrated a multi-scale version of our PBS framework.

All these steps were combined to create the volBrain pipeline. This pipeline was completed with an automatic report generation. The proposed framework is able to produce accurate segmentation for all considered structures in less than 15 minutes (see Results section of this chapter). Finally, to propose a tool robust, the training library of manually labeled templates was constructed using 50 subjects covering a wide range of age and including

pathological cases (see (Manjon and Coupe 2016) for details). This point is crucial to build reference models and pathological models across the entire lifespan using the same tool. We will demonstrate the robustness of our approach in the following. Once the volBrain pipeline finished, we used it to develop reference models across the entire the lifespan.

3.2.2 Lifespan Models

The study of normal brain maturation and age-related brain atrophy is crucial to better understand normal brain development and a large variety of neurological disorders. With the rise of the population age, it is becoming increasingly important to understand the cognitive changes that accompany aging, both normal and pathologic. Moreover, analyzing brain maturation and senescence during the entire lifespan may help to better understand the undergoing process on normal brain development and aging. Finally, the development of reference lifespan models is essential for performing brain monitoring that is the main goal of the proposed volBrain web-platform.

Despite the large number of studies dedicated to brain trajectory analysis, an important disagreement remains between existing results. An extensive review of these inconstancies can be found in Walhovd et al. (2016). This lack of consensus on brain development and aging prevents us to better understand these highly complex and multi-factor phenomena. The significant divergence between existing results is due to many factors:

- First, the use of restricted life periods (e.g., childhood (Brain Development Cooperative 2012), adolescence (Lenroot and Giedd 2010, Vijayakumar et al. 2016), adulthood (Ziegler et al. 2012), etc.) makes difficult the comparison of results. Moreover, it prevents global understanding of brain modification across the entire lifespan. Up to now, no study covered the entire lifespan including babies with few months of life to elderly older than 90.
- Second, the use of a limited number of scans may produce unstable results limiting the reproducibility and accuracy of estimations. The large majority of previous studies used less than 100 subjects (Walhovd et al. 2011), some studies used several hundreds of subjects (Giedd and Rapoport 2010, Brain Development Cooperative 2012, Ziegler et al. 2012, Mills et al. 2016) and very few studies used more than 1000 subjects (Fjell et al. 2013, Potvin et al. 2016).
- In addition, the use of non-harmonized acquisition protocols, segmentation tools, labelling protocols (Walhovd et al. 2016) and volumetric measurements such as absolute volume (Brain Development Cooperative 2012), normalized volumes using intracranial volume (Good et al. 2002, Mills et al. 2016), GM volume (Ziegler et al. 2012) or z-scores (Ostby et al. 2009, Walhovd et al. 2011), lead to a great discrepancy in reported results (Walhovd et al. 2011). Consequently, this heterogeneity makes difficult the definition of normative values (Potvin et al. 2016) stressing the need of using harmonized protocols over large samples covering the entire lifespan.
- Finally, the use of an exigent quality control in the whole measurement process plays a major role in the quality of the final estimated brain models. This step is often underconsidered, while the model estimation greatly depends on a careful quality control (Ducharme et al. 2016).

In (Coupe et al. 2017), to build our reference lifespan models, we have addressed these limiting factors. First, thanks to the new paradigm of Big Data sharing in neuroimaging (Poldrack and Gorgolewski 2014), we have been able to use a very high number of samples (N=3296) covering the largest lifespan period never studied (from few months to advanced ages). Moreover, all the considered MRI scans were processed using the same processing, our volBrain pipeline. To get insight on brain maturation and aging at global (i.e., absolute volume) and brain scale (i.e., normalized volume), we have extensively analyzed our results using absolute volumes and relative volumes (normalized by Total Intracranial Volume, TIV=ICC). Moreover, to prevent the estimated models to be affected by wrongly processed images, we have performed a demanding three stages quality control. Finally, to be able to present a unified analysis of brain development and brain aging at the same time, we investigated hybrid models. Contrary to previous studies based on linear or low order polynomial models, we considered models enable to capture fast growth and complex degenerative processes. This is achieved by combining cumulative exponential function – to model rapid growth with saturation related to brain maturation – and low order polynomial function – used to model volume decrease caused by aging. Putting all these elements together, we have been able to show for the first time a global picture of brain trajectory across the entire lifespan.

3.2.2.1 Data Description

To build our reference models, we used 3D T1w MRI obtained from nine freely available databases covering the entire lifespan (Coupe et al. 2017)**.** The summary of used databases is detailed in Table 15**.**

Table 15: Dataset description used for the reference models. This table provides the name of the dataset, the MR acquisition configuration, the number of considered image before and after QC, the gender proportion after QC and the average mean, standard deviation in parentheses and the interval in brackets.

DATASET	Acquisition	Before QC	After QC	Gender after QC	Age in years after QC
C-MIND	1 site with 3T MR scanner	266	236	F = 129 M =107	8.44 (4.35) [0.74-18.86]
NDAR	10 sites with 1.5T and 3T MR scanner	612	382	F = 174 M = 208	12.39 (5.94) [1.08-49.92]
ABIDE	20 sites with 3T MR scanner	528	492	F = 84 M = 408	17.53 (7.83) [6.50-52.20]
ICBM	1 sites with 1.5T MR scanner	308	294	F = 142 M = 152	33.75 (14.32) [18-80]
IXI	3 sites with 1.5T and 3T MR scanner	588	573	F = 321 M = 252	49.52 (16.70) [20.0- 86.2]
OASIS	1 sites with 1.5T MR scanner	315	298	F = 187 M = 111	45.34 (23.82) [18 - 94]
AIBL	2 sites with 1.5T and 3T MR scanners	236	233	F = 121 M =112	72.24 (6.73) [60 - 89]
ADNI 1	51 sites with 1.5T MR scanner	228	223	F = 108 M = 115	75.96 (5.03) [60 – 90]
ADNI 2	14 sites with 3T MR scanners	215	213	F = 113 M = 100	74.16 (6.39) [56.3 - 89]
Total	**103 sites with 1.5T and 3T scanners**	**3296**	**2944**	**F =1379 (47%)** **M = 1565 (53%)**	**39.65 (26.62)** **[0.74 - 94]**

The used images have been acquired on 1.5T and 3T over 103 sites. All the considered subjects are normal controls. After quality control, 2944 MRI were kept from the 3296 considered subjects. The gender proportion of these selected subjects is 47% of female. The covered age starts from 9 months to 94 years. Figure 25 shows the age distribution of the used subjects after quality. At least three different datasets are used for all the considered periods except for extreme ages (i.e., [0-4] year and [90-94] year) where only 2 datasets are available. Moreover, more than 50 subjects by 5-years interval are used at the exception of the last [90-94] interval.

As recently shown, the quality control (QC) of image processing pipeline has a critical impact on trajectory results (Ducharme et al. 2016). Therefore, to build our reference models, we decided to use a demanding multi-stage QC procedure for a careful selection of involved subjects. First, a visual assessment of input image quality was done for all considered subjects. This assessment was performed by checking screen shots of one sagittal, one coronal and one axial slice in middle of the 3D volume. This step led to remove 219 subjects from the 3296 considered subjects in our study (6.6%). Next, a visual assessment of the image processing quality for all remaining subjects was performed using volBrain reports (see Figure 33). All the volBrain pipeline steps (full head coverage including cerebellum, registration to MNI space, TIV extraction, tissue classification, subcortical structure segmentation, etc.) were carefully checked. This step led to remove 83 subjects from our study (2.5%). Finally, a last control was performed by individually checking all outliers detected using estimated trajectories. A volume was considered as outlier when its value was higher/lower than 2 standard deviations of the estimated model. For each detected outlier, the segmentation map was opened and displayed over the MRI using a 3D viewer (Yushkevich et al. 2006). A careful inspection was performed over the 3D volume. In case of segmentation failure, the subject was removed from the study. This last QC step led to remove 50 subjects (1.5%). Therefore, 2944 of the 3296 considered subjects were kept after our QC procedure.

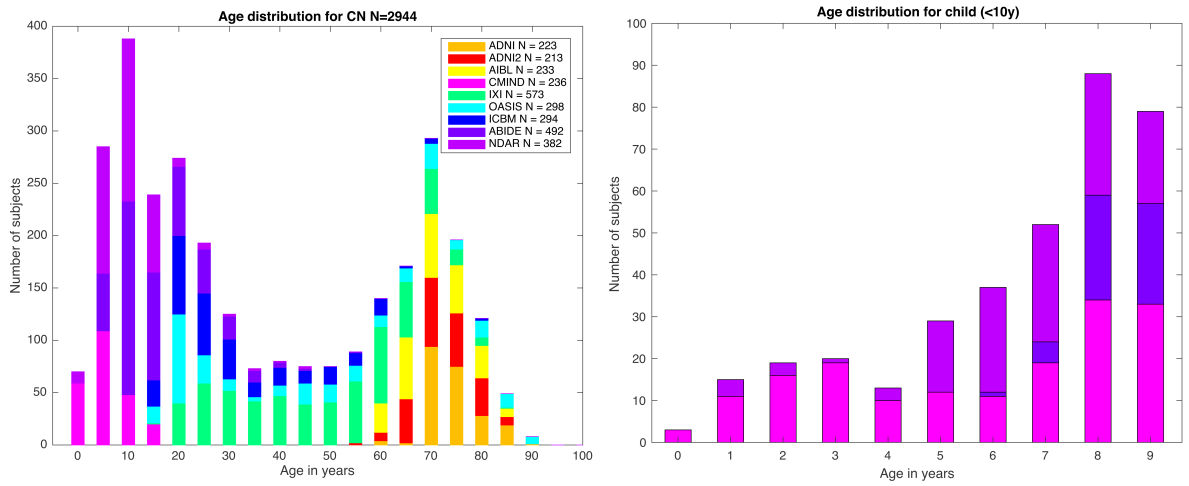

Figure 25: Age distribution of the used MRI after the quality control. Left: Age distribution for all the considered subjects. Right: Age distribution for child younger than 10 years old. Legend indicates the database color and the number of image after quality control.

In the following, we will also present our last work on pathological lifespan model dedicated to AD (Coupé et al. 2018). This study aimed at comparing normal and pathological trajectories of brain atrophy during AD progression across the entire lifespan to better understand pathology progression. In this work, we assumed that the neurodegenerative process is slow and progressive. Consequently, to build our lifespan pathological models we combined AD and MCI (from 55y to 96y) with healthy/asymptomatic subjects younger than them (from 9 months to 55y). The proposed approach can be viewed as a conservative lifespan model of AD since CN are used as young asymptomatic AD subjects. The summary of used databases to estimate our pathological models is presented in Table 16. Different models were estimated based on four different groups to generate CN, AD/MCI, AD and MCI trajectories.

- For CN trajectories, we used the N=2944 subjects from 9 months to 94y of the cognitively normal dataset described previously (see Table 15).
- For the AD/MCI trajectories, we used N=3262 samples. We mixed AD patients, MCI patients considered being at an early stage of AD (see Table 16), and young CN considered as presymptomatic subjects. We used 426 AD patients (from 55y to 96y), 959 MCI patients (from 55y to 92y) of the AD/MCI dataset (see Table 16) and all the CN younger than 55y (i.e., 1877 samples) described in Table 15.
- For the AD trajectories, we used N=2303 samples. We mixed AD patients with young CN. More precisely, we used 426 AD patients (from 55y to 96y) of the AD/MCI datasets and all the CN younger than 55y (i.e., 1877 samples).
- For the MCI trajectories, we used N=2836 samples. We mixed 959 MCI patients (from 55y to 92y) of the AD/MCI datasets with all the CN younger than 55y (i.e., 1877 samples)

All the MRI of AD and MCI patients and CN subjects followed the same processing and quality control than previously described. Therefore, at the end, 4714 MRI processed with the same tool were used during our pathological model study.

Table 16: Dataset description used for the AD models. This table provides the name of the dataset, the MR acquisition configuration, the number of considered image before and after quality control, the gender proportion after quality control and the average mean, standard deviation in parentheses and the interval in brackets.

DATASET	Before QC	After QC	AD stage (MCI /AD) after QC	Gender after QC	Age in years after QC
OASIS	98	95	50 / 45	F = 56 M = 39	76.58 (7.18) [62 - 96]
AIBL	112	106	59 / 47	F = 58 M =48	74.15 (7.80) [55 - 93]
ADNI 1	587	568	385 / 183	F = 225 M = 343	75.04 (7.41) [55 – 91]
ADNI 2	621	616	465 / 151	F = 270 M = 346	72.56 (7.64) [55 - 90]
Total	**1418**	**1385**	**959 / 426**	**F = 609 (44%)** **M = 776 (56%)**	**73.7 (7.84)** **[55 - 96]**

3.2.2.2 Statistical Analysis

In order to determine the best general models for each structure, several models were tested from the simplest to the most complex. A model was kept as a potential candidate only when F-statistic based on ANOVA for model vs. constant model is significant ($p<0.05$) and when all its coefficients were significant using t-statistic ($p<0.05$). At the end of the selection procedure, we used the Bayesian Information Criterion (BIC) to select the best model among models being significant compared to constant model and having all coefficients significant. BIC provides a measure of the trade-off between bias and variance and thus select the model explaining most the data with minimum parameters. Afterwards, this general model type is applied on female and male separately to estimate gender specific reference models. At the end, to study trajectory difference in terms of volume and shape between both female and male, $\beta_i Sex + \beta_j Sex.Age$ interactions are tested over the selected general model. All the reported parameters (t-statistic, F-statistic, BIC and R^2) were internally estimated by Matlab© using default parameters. The following models were considered as potential candidates:

1. Linear model

$$Vol = \beta_0 + \beta_1 Age + \varepsilon \quad \text{(Eq. 10)}$$

2. Quadratic model

$$Vol = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \varepsilon \quad \text{(Eq. 11)}$$

3. Cubic model

$$Vol = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Age^3 + \varepsilon \quad \text{(Eq. 12)}$$

4. Linear hybrid model: exponential cumulative distribution for growth with linear model for aging

$$Vol = \beta_4 . \left(1 - e^{-Age/\beta_5}\right) + \beta_0 + \beta_1 Age + \varepsilon \quad \text{(Eq. 13)}$$

5. Quadratic hybrid model: exponential cumulative distribution for growth with quadratic model for aging

$$Vol = \beta_4 . \left(1 - e^{-Age/\beta_5}\right) + \beta_0 + \beta_1 Age + \beta_2 Age^2 + \varepsilon \quad \text{(Eq. 14)}$$

6. Cubic hybrid model: exponential cumulative distribution for growth with cubic model for aging

$$Vol = \beta_4 . \left(1 - e^{-Age/\beta_5}\right) + \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 Age^3 + \varepsilon \quad \text{(Eq. 15)}$$

In the literature, structure trajectories have been mainly modeled using low order polynomial function (see (Walhovd et al. 2011) for review). However, to follow structure trajectories across the entire lifespan, we propose to consider hybrid models able to track rapid growth during childhood and to capture complex volume decrease from adulthood to elderly. In the past, fast growth modelling occurring during childhood has been achieved using Poisson curve (Lebel et al. 2012) or Gompertz-like function (Makropoulos et al. 2016). Here, we propose to combine a cumulative exponential function in place of Gompertz-like function, and to combine it with low order polynomial function. At the end, our hybrid models can model fast growth process and complex volume decreases at the same time.

3.3 Results

3.3.1 Performance of the volBrain Pipeline

In this subsection, some experimental results are shown to highlight the accuracy and reproducibility of the proposed volBrain pipeline. Since volBrain provides results at different scales both accuracy and reproducibility at several scales will be commented. Specifically, we will comment the results for intracranial cavity extraction (NICE), macrostructure segmentation (NABS) and subcortical structure segmentation. Note that the tissue classification is not included in this evaluation since it is based on our particular way to compute PVCs. Therefore, there is no a direct comparison with methods like SPM or VBM.

3.3.1.1 Intracranial Cavity Extraction

NICE results for ICC extraction accuracy were presented in its corresponding paper (Manjon et al. 2014). To summarize, NICE was compared with BEaST and VBM8 and it was found to be significantly better. NICE obtained the best DICE coefficient (0.9911) compared to BEaST (0.9880) and VBM8 (0.9762). Moreover, an independent test was also performed using the SVE website (see http://sve.bmap.ucla.edu/archivel/) were NICE ranked first. Regarding the reproducibility, NICE was found to be the most reproducible method followed by VBM8 and finally BEaST. More details can be found the original paper (Manjon et al. 2014).

3.3.1.2 Hemisphere Segmentation

NABS method was also evaluated in its corresponding paper for hemisphere segmentation (Romero et al. 2015). First, NABS was compared with ADISC (Zhao et al. 2010) and it obtained a significantly better DICE coefficient (0.9962 vs. 0.9868 for ADISC). NABS method was also compared to ADISC method using an application consisting on estimating brain asymmetries on AD cases. We showed that NABS method was able to better predict the patient status. Again, further details can be found in the original paper (Romero et al. 2015).

3.3.1.3 Structure Segmentation

Finally, we propose experiments to measure both accuracy and reproducibility of the proposed subcortical segmentation method and a comparison with *state-of-the-art* approaches.

- Accuracy

First, a leave-two-out cross validation procedure was performed for the 50 subjects of the volBrain training library. DICE value was then computed by comparing the manual segmentations with the segmentations obtained with our method. The proposed method was also compared with two publically available software packages for subcortical brain structures labeling (Freesurfer (Fischl 2012) and FSL-FIRST (Jenkinson et al. 2012). Both methods were run on the CBRAIN platform with their default parameters. As can be noted on Table 17, the volBrain pipeline obtained the best DICE coefficients for all the considered structures.

*Table 17: Mean Dice coefficient of the different subcortical structures over the 50 cases of template library. Average represents the average dice without including lateral ventricles. Best results are in bold. * indicates statistically significant differences between volBrain and Freesurfer ($p<0.05$). † indicates statistically significant differences between volBrain and FIRST.*

Structure	volBrain	Freesurfer	FIRST
Lat. Ventricles	**0.9836 ± 0.0111 ***	0.8315 ± 0.0589	---
Caudate	**0.9427 ± 0.0196 *†**	0.8195 ± 0.0418	0.8366 ± 0.0706
Putamen	**0.9442 ± 0.0226 *†**	0.8162 ± 0.0396	0.8775 ± 0.0192
Thalamus	**0.9476 ± 0.0196 *†**	0.8157 ± 0.0247	0.8144 ± 0.0322
Pallidum	**0.8914 ± 0.0403 *†**	0.7454 ± 0.0906	0.7851 ± 0.0575
Hippocampus	**0.9533 ± 0.0092 *†**	0.7886 ± 0.0254	0.8429 ± 0.0303
Amygdala	**0.8795 ± 0.0559 *†**	0.5844 ± 0.1092	0.5895 ± 0.0962
Accumbens	**0.8362 ± 0.0572 *†**	0.5589 ± 0.0697	0.6483 ± 0.0916
Average	**0.9136 ± 0.0555 *†**	0.7327 ± 0.1132	0.7706 ± 0.1087

Moreover, the improvement was statistically significant for all the structures and for the two methods compared. The volBrain pipeline obtained an average dice coefficient (without including lateral ventricles) of 0.9136 while Freesurfer obtained 0.7327 and FIRST obtained 0.7706. A visual comparison of one sample case is showed Figure 26 were the labeling of the three different methods are presented with 3D representation (note that FIRST does not segment lateral ventricles and therefore they are not included in the comparative). On one hand, Freesurfer method produced a rough segmentation of the different structures with significant errors. On the other hand, FIRST performed better and produced smoother surfaces on the different structures. However, FIRST method seems to over segment most of the structures.

We are aware that the presented accuracy results are slightly biased in favor to volBrain due to the use of the same label definition for training and validation. However, there are minimal differences on our label definition compared to FIRST or Freesurfer labels with the exception of lateral ventricles (we did not include choroid plexus) and HIPP (we used EADC protocol). Besides, it has been recently shown that Freesurfer and FIRST overestimate most of subcortical structures by using a joint DTI-MRI analysis (Næss-Schmidt et al. 2016) that matches with our findings. In summary, the large differences found among the compared methods provide evidences of the high quality of the proposed pipeline. Moreover, it is important to note that volBrain is one of the first software following the harmonized EADC protocol for HIPP segmentation. This is especially important given that fact that EADC protocol is the new consensus protocol for HIPP analysis dedicated to AD (Frisoni et al. 2015).

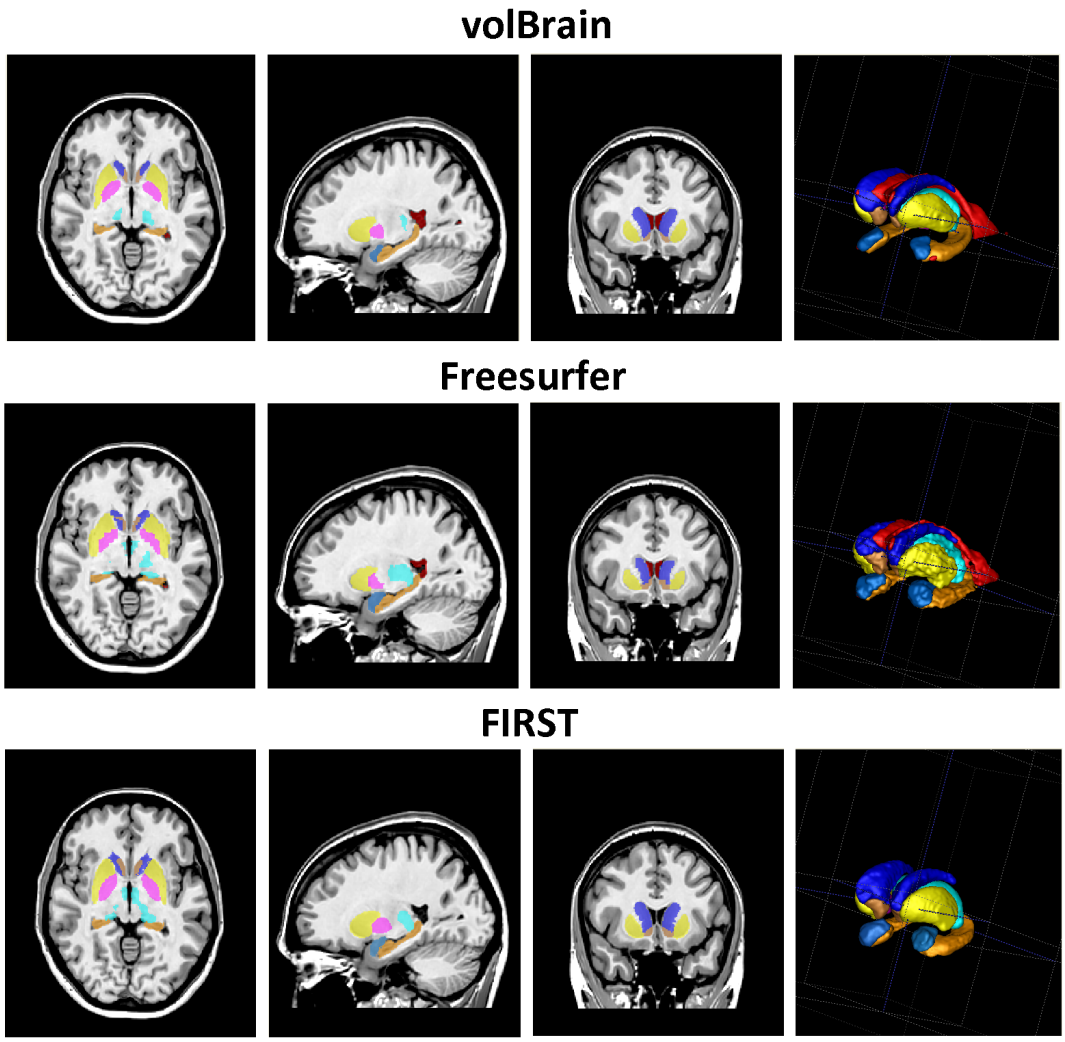

Figure 26: Visual example of the segmentation results (Axial, sagittal and coronal views and 3D representation of segmented subcortical structures). First row: volBrain results. Second row: Freesurfer results. Third row: FIRST results. Note that FIRST output does not include lateral ventricles.

- Reproducibility

Second, we investigated the reproducibility of our volBrain pipeline. To measure the reproducibility of the different methods compared, we used the reproducibility dataset of the brain segmentation testing protocol website (https://sites.google.com/site/brainseg/). This dataset consists of a test-retest set of 20 subjects scanned twice in the same scanner and with the same sequence. To measure the reproducibility of the two repeated sets, we used the Percent Volume Difference (PVD) (Morey et al. 2010). The three compared methods were run on this dataset but the comparison was done only on a subset of 18 subjects since FIRST did not produce valid results for two of the 20 cases. Since PVD measurement does not distribute normally, we represent the results using the median and the interquartile interval, and we used the Wilcoxon rank test to measure the statistical significance of the differences between methods. Results of this comparison are summarized on Table 18. As can be noted, volBrain was more reproducible in general compared to Freesurfer and FIRST (although the differences were not statistically significant overall). Regarding to the volume estimation, volBrain was found to be significantly more reproducible than Freesurfer for putamen ($p<0.05$) while FIRST was significantly better than volBrain and Freesurfer for the pallidum.

*Table 18: Median of PVD on the reproducibility dataset and corresponding interquartile interval (shown in brackets). Best results are in bold. * indicates statistically significant differences between volBrain and Freesurfer ($p<0.05$). † indicates statistically significant differences between volBrain and FIRST. Overall represents the mean PVD of all structures excluding lateral ventricles to enable direct global comparison of the three methods.*

Structure	volBrain	Freesurfer	FIRST
Lat. Ventricles	**4.95 [7.42]**	5.06 [4.85]	--
Caudate	**0.53 [1.22]**	1.39 [2.47]	1.19 [1.20]
Putamen	**0.69 [1.76] ***	2.23 [2.90]	1.20 [1.23]
Thalamus	**0.82 [1.04]**	0.93 [0.66]	0.98 [1.67]
Pallidum	1.60 [1.92]	3.18 [3.72]	**0.92 [0.96] †**
Hippocampus	**1.41 [2.97]**	1.73 [1.74]	2.15 [3.20]
Amygdala	**3.38 [2.06]**	4.13 [5.23]	3.94 [3.63]
Accumbens	**2.65 [22.81]**	2.68 [2.48]	4.26 [6.92]
Overall	**1.59**	2.33	2.09

- Computational Time

The proposed method takes around 12 minutes in average to complete the whole pipeline, this included 30 seconds for denoising, 2 minutes to perform registration into MNI space, 3 minutes for inhomogeneity correction, 5 seconds for intensity normalization, 2 minutes to do brain extraction, 5 seconds to perform tissue classification, 2 minutes for NABS and 3 minutes for structures labeling. Freesurfer normally takes around 15 hours to perform the complete analysis (which includes also surface extraction). FIRST running time is approximately 10 minutes (only for the subcortical structure segmentation).

3.3.2 Analysis of Lifespan Reference Models

Once the volBrain pipeline validated, we used it to process a massive number of MRI to build reference models. In this subsection, we will analyze the estimated models and the new produced neuroscientific knowledge. We will present results in absolute volumes and relative volumes in % of Total Intracranial Volume (i.e., ICC).

3.3.2.1 Global Gray Matter and White Matter Trajectories

At the global scale (i.e., absolute volumes in native space), we observed an increase of WM volume until 30-40y followed by a volume decrease (see Figure 27). The WM growth at early ages is faster than the senescence at late ages. This is assessed by the selected hybrid model ($p<0.001$) combining an exponential cumulative distribution model for growth and a cubic model for aging (see Eq. 15). On the other hand, although the same model is selected for GM ($p<0.0001$), its trajectory is more complex. We can observe a 4-stage trajectory composed of a fast growth until 8-10y followed by a fast decrease until 40ys, then a plateau and finally an accelerated aging-related decrease visible around 80ys.

At the brain scale, when using normalized volumes in % to the TIV, the main difference is found for the GM trajectory. Indeed, we observed a decrease of GM all along the lifespan (see Figure 28) following a cubic model ($p<0.0001$). The decrease of normalized volumes also follows a complex shape with 3 stages composed of a rapid decrease from 0 to 20y, a plateau from 40 to 80y and a rapid decrease after 80y. It is interesting to note that despite the normalization, the WM growth remains very fast at the brain scale for early age since a hybrid model using an exponential cumulative distribution model for growth has been selected.

Finally, at global and brain scales, we observe that WM have almost an inverted U-shape model although an asymmetry exists with a faster volume increase related to maturation than volume decrease caused by aging.

3.3.2.2 Deep Gray Matter Structure Trajectories

3.3.2.2.1 Thalamus, Accumbens, Caudate, Putamen and Globus Pallidus Trajectories

At global scale, we observed that thalamus, accumbens, caudate and putamen follow similar trajectories with fast growth until 10-12y followed by a volume decrease. All selected hybrid models combine an exponential cumulative distribution for growth followed by low polynomial order for volume loss during aging, cubic for caudate ($p<0.0001$) and putamen ($p<0.0001$), quadratic for thalamus ($p<0.0001$) and linear for accumbens ($p<0.0001$). On the other hand, globus pallidus volume decreases from birth all along lifespan. Unexpected slight increases of caudate and putamen volumes are visible after 80y.

At the brain scale, we observed that thalamus, accumbens, caudate, putamen and globus pallidus show a volume decrease across the entire lifespan. First, thalamus and accumbens exhibit almost monotonous decrease although cubic models have been selected (both with $p<0.0001$). Second, caudate and putamen present similar slowdown decreases after 50y. The similar trajectories of the caudate and putamen are consistent with their shared nature as dorsal striatal structures (Paxinos and Mai 2004). The model selected for these structures is cubic for caudate ($p<0.0001$) and quadratic for putamen ($p<0.0001$). Finally, globus pallidus follows a cubic model ($p<0.0001$).

3.3.2.2.2 Amygdala and Hippocampus Trajectories

At the global scale, amygdala volume shows a slight increase until 18y-20y followed by a long plateau that ends around 70y, followed by an age-related atrophy. The selected hybrid model combines a volume increase following an exponential cumulative distribution and a volume decrease following cubic model ($p<0.0001$). The hippocampus trajectory presents a fast volume increase until 8y-10y followed by a slow volume increase until 40y-50y before an atrophic period. Here, the selected hybrid model mixes a volume increase following an exponential cumulative distribution and then an inverted U-shape volume decrease ($p<0.0001$).

At the brain scale, amygdala volume trajectory follows a cubic model ($p<0.0001$) with a plateau until 70y followed by an atrophy. This result seems to indicate that absolute increase of amygdala volume during childhood is mainly related to brain growth. Moreover, using relative volume, HIPP exhibits a very specific inverted U-shape trajectory compared to other analyzed subcortical structures. In our study, the HIPP was the only structure showing volume increase until the middle period of human life. To better investigate this point, we performed a complementary analysis between 18y and 70y. We found that the impact of age on absolute HIPP volume is significant ($p<0.0001$) and that the selected model is an inverted U-shape trajectory over this restricted period. According to our results, the hippocampal maturation stops around 50y.

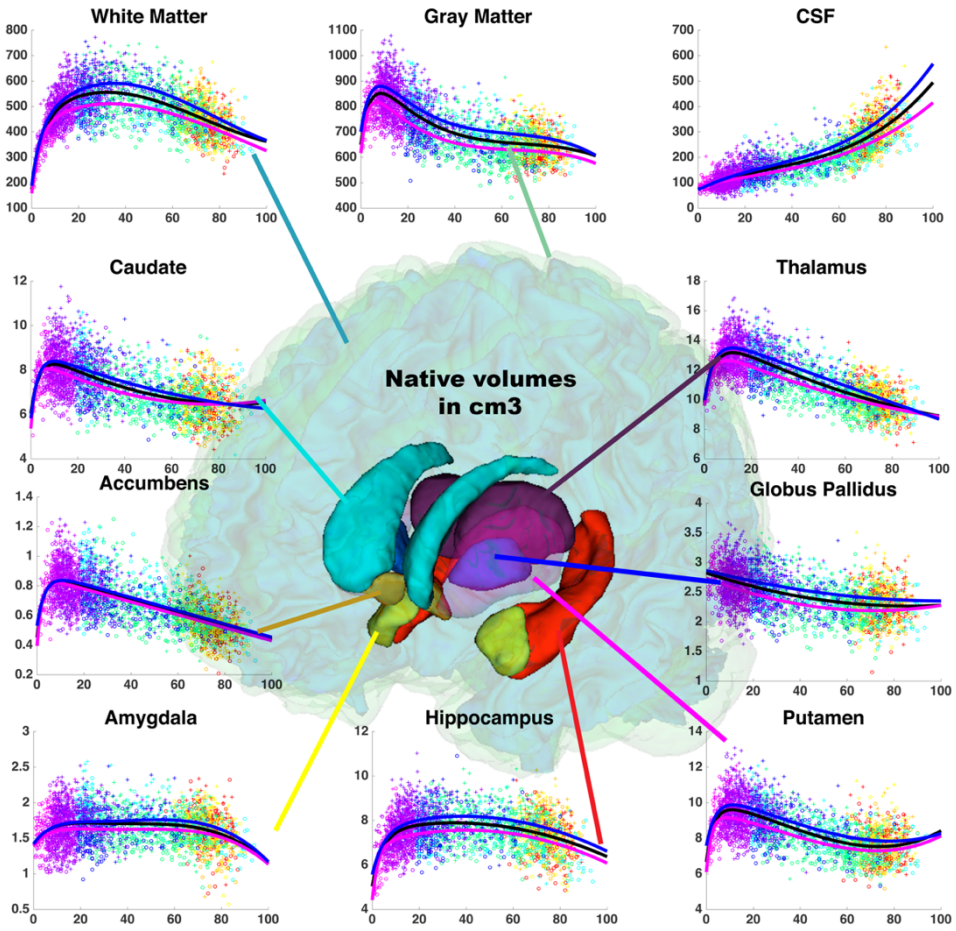

Figure 27: Volume trajectories based on absolute volume in cm3 for brain tissues and subcortical structures across the entire lifespan. These volume trajectories are estimated according to the age on 2944 subjects from 9 months to 94 years. General model is in black, female model is in magenta and male model is in blue. Dots color represents the different datasets used in this study.

3.3.2.3 Sexual Dimorphism

At the global scale, we observe that males have bigger volumes than females for all considered structures (sex interaction with $p<0.0001$) with the exception of accumbens. Moreover, the peaks of maturation appear 1-3 years earlier in females for most of the structures considered. Finally, increased atrophy rates for males after 80y is assessed by CSF trajectory, which is the only brain compartment showing significant age*sex ($p<0.0001$).

At the brain scale, almost all gender volume differences vanish, except for caudate ($p=0.05$) and thalamus ($p=0.05$) with marginal significance, and for accumbens ($p=0.02$) all in favor of females. Visually, we can observe bigger relative HIPP volume for female (almost significant $p=0.07$) (see Figure 28). Finally, for global GM, caudate, thalamus, globus pallidus and amygdala, trajectories of females seem to indicate a better resistance to the accelerated age-related atrophy occurring after 80y. To investigate this point, we studied sex and sex*age interaction using all subjects with age > 70 years (i.e., 637 subjects composed of 292 males and 345 females). Models estimated using all the subjects are applied over this considered restricted period to evaluate sex and sex*age interactions. We found that using normalized volumes, almost all studied structures show significant sex and sex*age interaction after 70y with the exception of WM and amygdala.

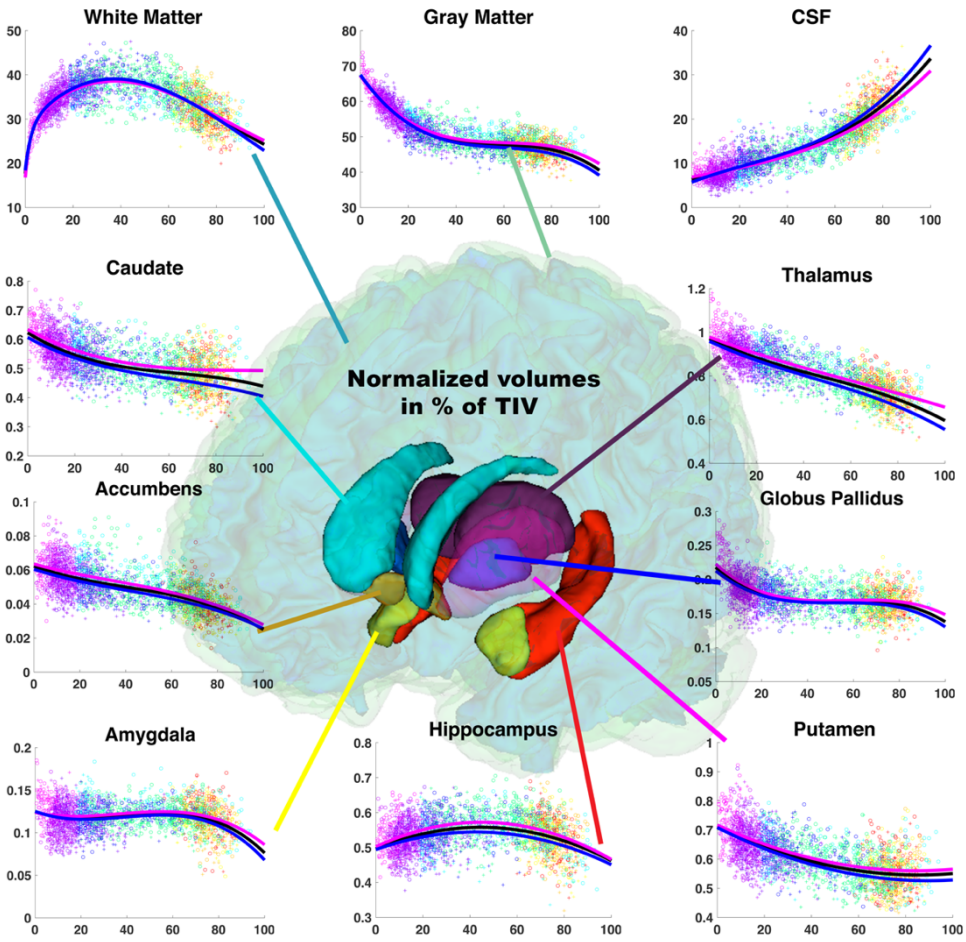

Figure 28: Trajectories based on relative volumes (% total intracranial volume) for brain tissues and subcortical structures across the entire lifespan. These volume trajectories are estimated according to the age on 2944 subjects from 9 months to 94 years. General model is in black, female model is in magenta and male model is in blue. Dots color represents the different datasets used in this study.

With our volBrain pipeline and our reference lifespan models, we have the two required elements to perform brain monitoring – a measurement tools and a range of reference (see Figure 23). It is the main realization of all the work described in Chapter 1 and 3. This illustrates the coherence and evolution of my work over the last 8 years. In the next subsection, we will compare our reference models with pathological models to demonstrate the usefulness of brain monitoring and to show how Big Data in neuroimaging can produce new medical knowledge.

3.3.3 Timeline of Brain Alterations in Alzheimer's Disease

Brain imaging studies have shown that progressive cerebral atrophy characterized the development of AD. The key question is – how long before the diagnosis of AD the neurodegenerative process started? To answer this question, in (Coupé et al. 2018), we proposed an innovative way by inferring brain structure volume trajectories across the entire lifespan using massive number of MRI using. As previously described, three pathological models were investigated – an AD model composed of 2303 samples, an MCI model composed of 2836 samples and an AD/MCI model composed of 3262 samples.

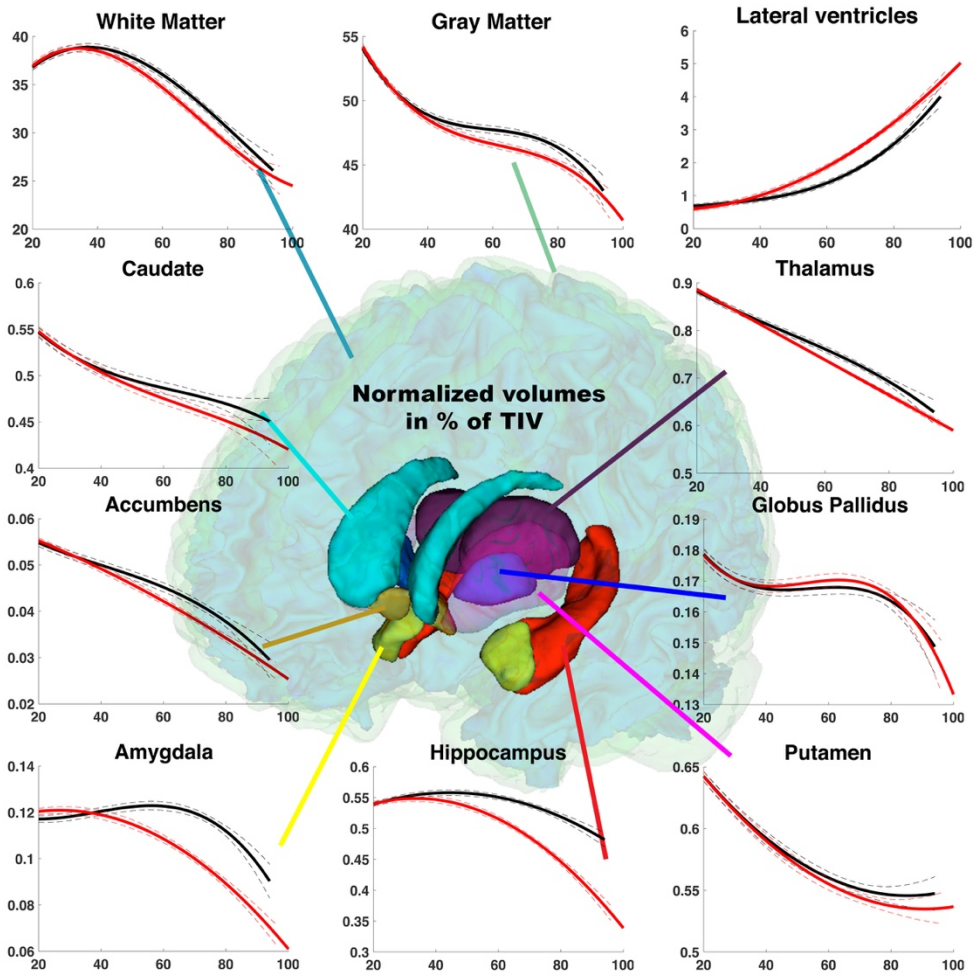

Figure 29: Trajectories based on relative volumes (% total intracranial volume) for brain cortical and subcortical structures across the entire lifespan. These volume trajectories are estimated according to the age of subjects. CN model (N=2944) is in black and AD/MCI model (N=3262) is in red. The prediction bounds are estimated with a confidence level at 95%.

Figure 29 presents trajectories of all considered structures for AD/MCI and CN groups. This figure shows that HIPP and amygdala models present marked divergences between AD/MCI and CN, but also indicates that this divergence increases with age. Moreover, the divergence of control and pathological models for these structures occurs early around 40-45y. Lateral ventricles also exhibit early divergence – starting around 42y –between both models, however the distance between models decreases at advanced ages. Similarly, the thalamus presents an early but weak divergence that decreases at advanced ages. Pathological models of caudate and accumbens nuclei exhibit accelerated volume decreases from 50-60y. However, confidence intervals for these structures overlap again after 85y (see Table 19). For WM and GM, AD/MCI models present an early accelerated aging compared to CN models around 45y. However, after 80y, CN models of brain tissues show an accelerated volume decreases. Consequently, confidence intervals of pathological and normal models for WM and GM overlap after 85 years (see Table 19). Finally, normal and pathological models for globus pallidus and putamen present similar trends. Table 19 shows the age ranges where the confidence interval of the 3 predicted pathological models (i.e., AD, MCI and AD/MCI) do not overlap with the confidence interval of the control models.

Table 19: Age range in years where confidence intervals of the predicted pathological models do not overlap with the predicted control models. The prediction bounds are estimated with a confidence level at 95%. Three model comparisons are presented CN (N=2944) vs. AD/MCI (N=3262), CN (N=2944) vs. AD (N=2303) and CN (N=2944) vs. MCI (N=2836).

	CN vs. AD/MCI	CN vs. AD	CN vs. MCI
White Matter	[47.6 - 85.8]	[46.9 – 89.9]	[53.7 - 82.3]
Gray Matter	[45.0 – 85.6]	[46.2 – 86.4]	[58.3 – 86.7]
Lateral Ventricles	[42.0 – 93.2]	**>38.6**	[45.1 – 89.2]
Caudate	[62.7 - 84.1]	[68.8 - 82.8]	[70.3 - 84.7]
Putamen	N/A	N/A	N/A
Thalamus	[42.8 – 89.1]	[41.7 – 89.6]	[45.5 – 86.7]
Globus Pallidus	N/A	N/A	N/A
Hippocampus	**>39.0**	**>37.1**	**>42.4**
Amygdala	**>43.8**	**>40.2**	**>49.3**
Accumbens	[48.1 – 85.6]	[46.0 – 88.0]	[52.6 – 82.3]

First, only HIPP and amygdala trajectories present non-overlapping confidence intervals after trajectory divergence for all the studied pathological modes (i.e., AD/MCI, AD and MCI). This is also valid for lateral ventricles but only when using the AD model. For all other considered structures, predicted confidence intervals overlap again at advanced ages around 80-90y.

Second, HIPP is the first impacted deep gray structure with a trajectory divergence at 39y when using the AD/MCI group, 37y when using the AD group and 42y for the MCI group. The second structure impacted is the lateral ventricles with a divergence point at 42y when using AD/MCI group, 39y for AD group and 45y for MCI group. Afterwards, thalamus trajectory divergence from control at 43y when using AD/MCI group, 42y for AD group and 45y for MCI group. Amygdala trajectory divergence occurs then at 44y when using AD/MCI group, 41y when using AD group and 49y for MCI group. Impact on global GM and WM volume is observed later, with trajectories diverging at 45y and 48y respectively for the AD/MCI group, at 46y and 47y respectively for the AD group and at 58y and 54y respectively for the MCI group. Finally, accumbens and caudate trajectories diverge slightly later, but in a similar age range. Putamen and globus pallidus are the only deep gray matter structures for which trajectories do not diverge from CN across the entire lifespan.

To further analyze trajectories of well-known AD biomarkers, we propose an additional analysis focusing on the HIPP, the lateral ventricles and the amygdala. Figure 30 presents the trajectories of these structures for CN, AD and MCI models. Moreover, relative rate of change and abnormality percentages are provided. First, divergence points for the AD models compared to the CN models are earlier than for the MCI models (see Table 19 for exact time). As expected, MCI trajectories are between AD and CN ones. Second, when using relative rate of change, amygdala and lateral ventricles exhibit a more pronounced relative change compared to HIPP. The maximum relative changes for AD models for these structures are -3.6%/y for amygdala at 96y, -2.1%/y for HIPP at 96y and 3.4%/y at 42y for lateral ventricles. Contrary to HIPP and amygdala, which show an increasing relative change with age, lateral ventricles exhibit enlargement following an inverted U-shape.

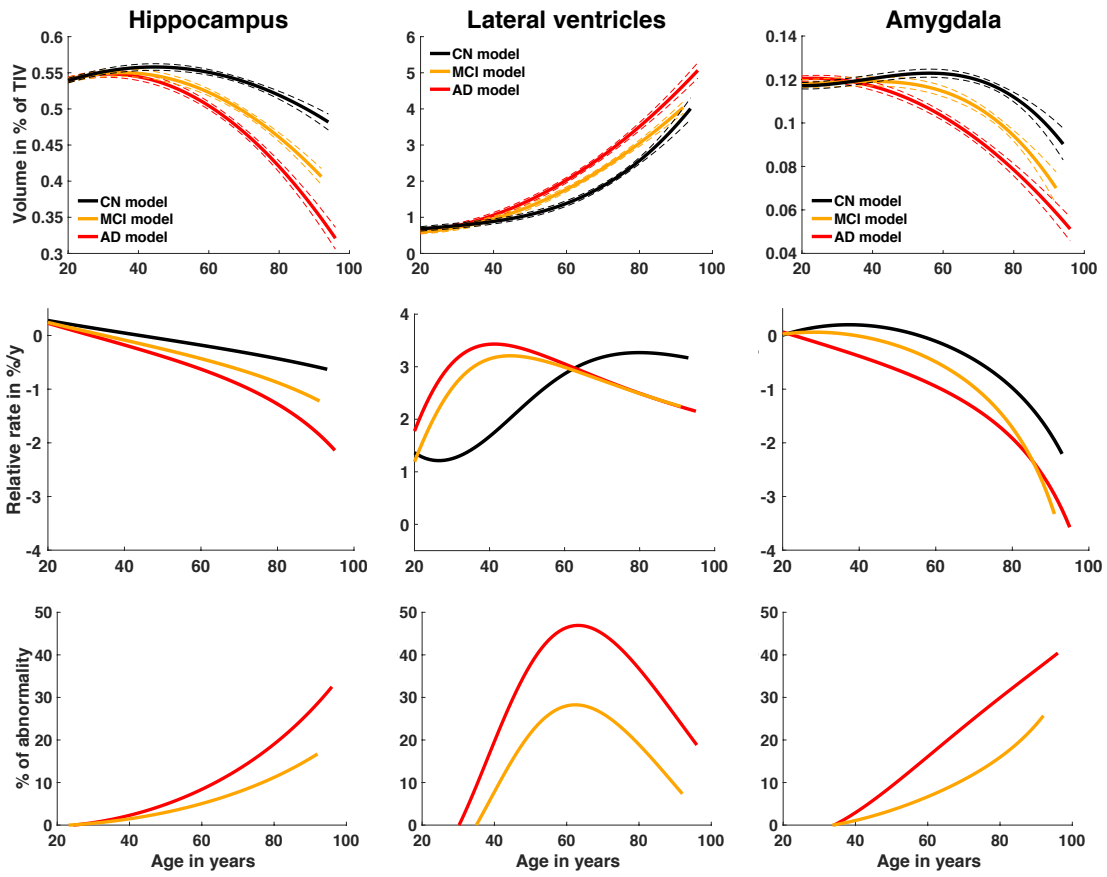

Figure 30: Hippocampus, lateral ventricles and amygdala trajectories for CN, AD and MCI models. The relative volumes (% total intracranial volume) are displayed according to the age in years across the entire lifespan. The prediction bounds are estimated with a confidence level at 95%. Relative rate of change is based on the first derivative of the model divided by the model and provided in % per year. Finally, percentage of abnormality is estimated as the difference between CN model and AD or MCI models divided by CN model. The model for CN group (N=2944) is displayed in black, the model for MCI group (N=2836) is displayed in yellow and the model for AD group (N=2303) is displayed in red.

When considering abnormality percentage, an earlier abnormality increase is observed for HIPP than for lateral ventricles and amygdala. This abnormality reaches a maximum of 32% for the AD model at 96y. Abnormality appears later in life for lateral ventricles and amygdala and follows very different patterns for both. The lateral ventricles abnormality trajectory follows an inverted U-shape with a maximum of 47% at 63y for the AD model. The amygdala abnormality trajectory is similar to the HIPP abnormality. Amygdala reaches 40% of abnormality at 96y for the AD model. Therefore, while HIPP abnormality starts first, amygdala presents a greater abnormality at advanced ages. Moreover, the abnormality observed in lateral ventricles is also important but its maximum is reached at 65y. Afterwards, percentage of abnormality of lateral ventricles decreases to end at 19% at 96y for the AD model. Therefore, at late ages, the lateral ventricles show lower abnormalities than those of the HIPP and the amygdala at the same age.

3.3.4 The volBrain Web-platform: 3 years of experience

Once the volBrain pipeline was finished and the reference lifespan models built, we integrated them into a fully open access web platform – the volBrain platform (http://volbrain.upv.es). In this subsection, we will briefly present the volBrain platform and its usage after 3 years of experience.

Most of the developed pipelines for MR image analysis are packages that need to be downloaded, installed and configured. Installation step can be complicated and thus may require an experimented person not always available in a research laboratory or clinical context. In addition, users have to be trained to use the software and computational resources have to be allocated to run it. These requirements can make complex the use of these packages, especially the most recent and sophisticated ones since they usually require advanced hardware configuration. Furthermore, multiplatform versions and support has to be deployed to the community of users. We have tried to overcome all these problems by deploying our volBrain pipeline through a web interface providing not only access to the software but also sharing the computational resources of our institutions. Thus, using the volBrain pipeline does not require any installation, configuration or training. The volBrain platform works remotely through a web interface based on a SaaS (Software as a Service) model to automatically provide a report containing volumetric information.

The volBrain pipeline is executed on dedicated clusters. The system has been designed to deal with up to 14 concurrent volBrain jobs and has a theoretical limit of 1200 processed cases per day. To get access to the system, the user has to register by providing personal information such as email, name and institution name. The web server (see Figure 31) accepts requests (jobs) from users who submit a single anonymized compressed MRI in Nifti file format *via* a web interface (see Figure 32). Then, the web server dispatches the computational load among the available machines called workers. This job dispatching is done by using a queuing system based on a job manager (technical details can be found in (Manjon and Coupe 2016)).

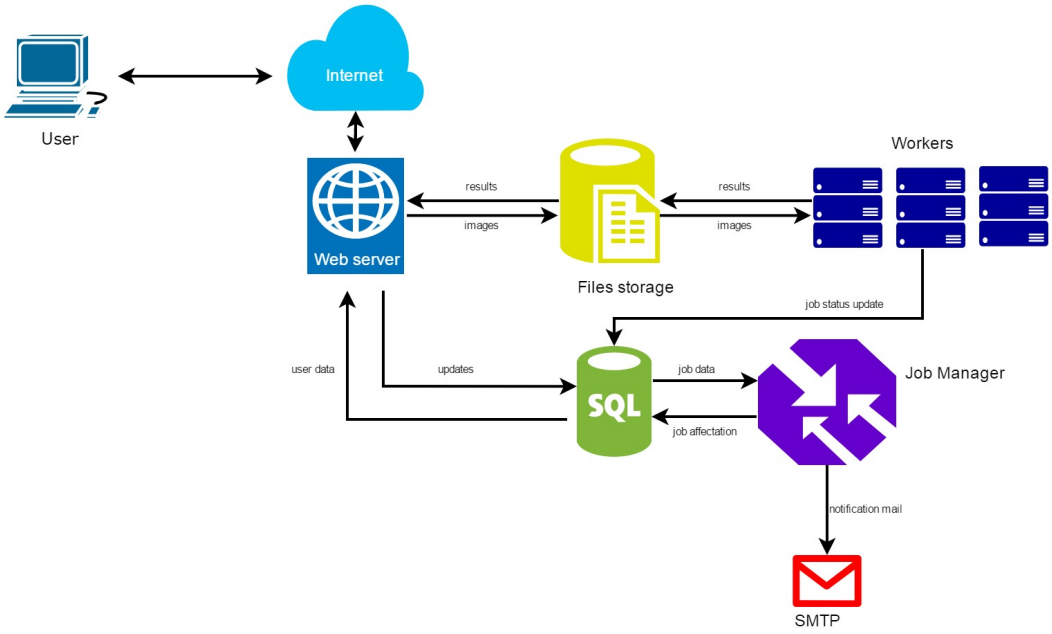

Figure 31: The volBrain platform architecture.

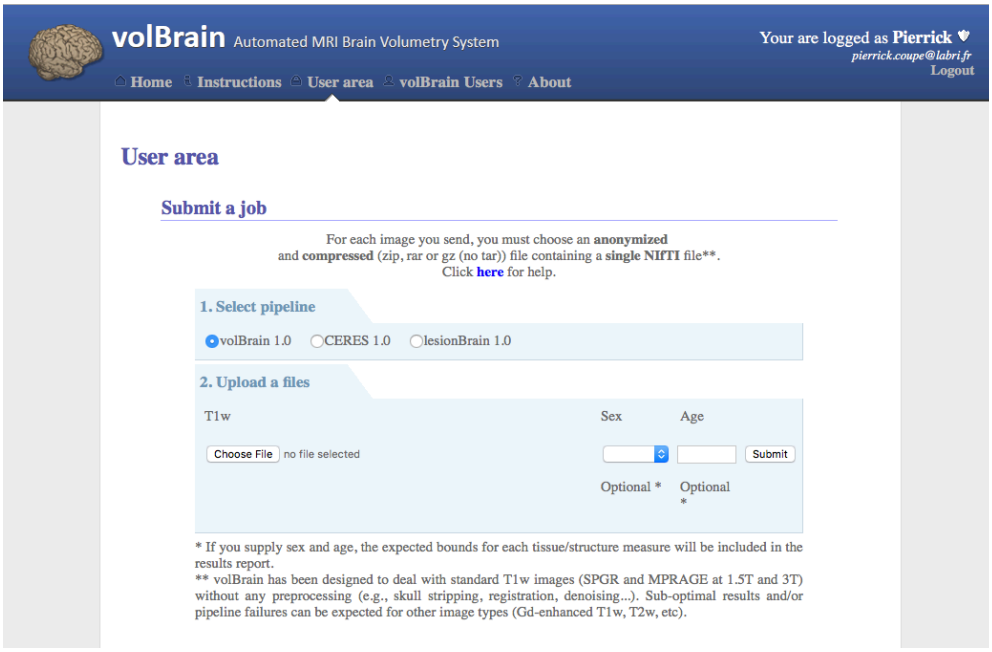

Figure 32: The volBrain website.

The outputs produced by the volBrain platform consists in report sent to the user by email using a SMPT server. This report summarizes the volumes and asymmetry ratios estimated from the submitted cases. If the user provides the sex and age of the uploaded subject, we compared the estimated volume to our reference models by providing normal bounds in the report. Furthermore, the user can download the resulting Nifti files containing the segmentations at different scales (in native or MNI space) in his user area directly from the volBrain website. Figure 33 shows an example of a report produced by the volBrain platform. Screenshots for each step of the pipeline are included for quality control purpose.

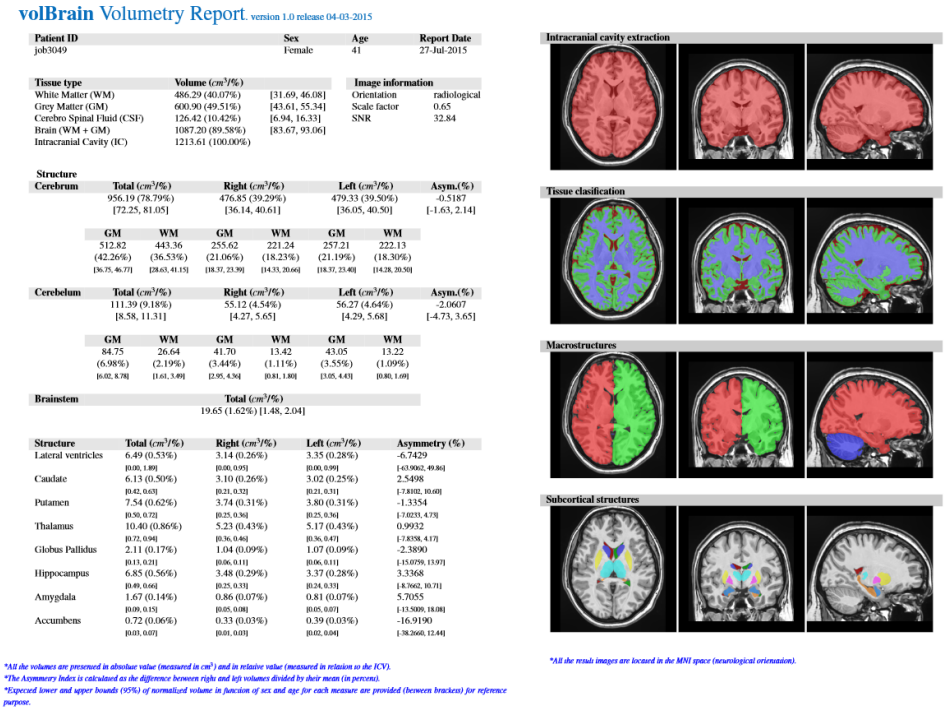

Figure 33: Example of volBrain pdf report.

The volBrain web platform has been launch officially the 28th Match 2015 after 6 months of beta testing. At the day of writing this manuscript, the system had already more than 1800 registered users working in 800 different institutions all around the world. Moreover, our system has automatically processed more than 70.000 subjects with a failure rate around 2%. Figure 34 shows a world map of the distribution of the volBrain users. Thanks to our fully open access philosophy, we can see that people all around the world are using our platform. At present, we have only one computational site with 7 servers in Spain. However, we are currently moving to a more modular architecture based on Docker technology (https://www.docker.com) to prepare the deployment of a second site at the LaBRI UMR 5800 of the University of Bordeaux. We have just installed the new servers in France. We are now working to manage both sites efficiently and synchronously.

Since its introduction, we added new pipelines to the volBrain platform. First, our cerebellum lobule segmentation method CERES (Romero et al. 2017a) presented in Chapter 2 is already proposed through our web-platform and account for 15000 jobs on the 70.000. Moreover, a new pipeline for white matter hyperintensity segmentation called lesionBrain (Guizard et al. 2015a) has been launch one week ago and already processed 500 cases. Finally, our hippocampus subfields segmentation method HIPS (Romero et al. 2016) is under integration.

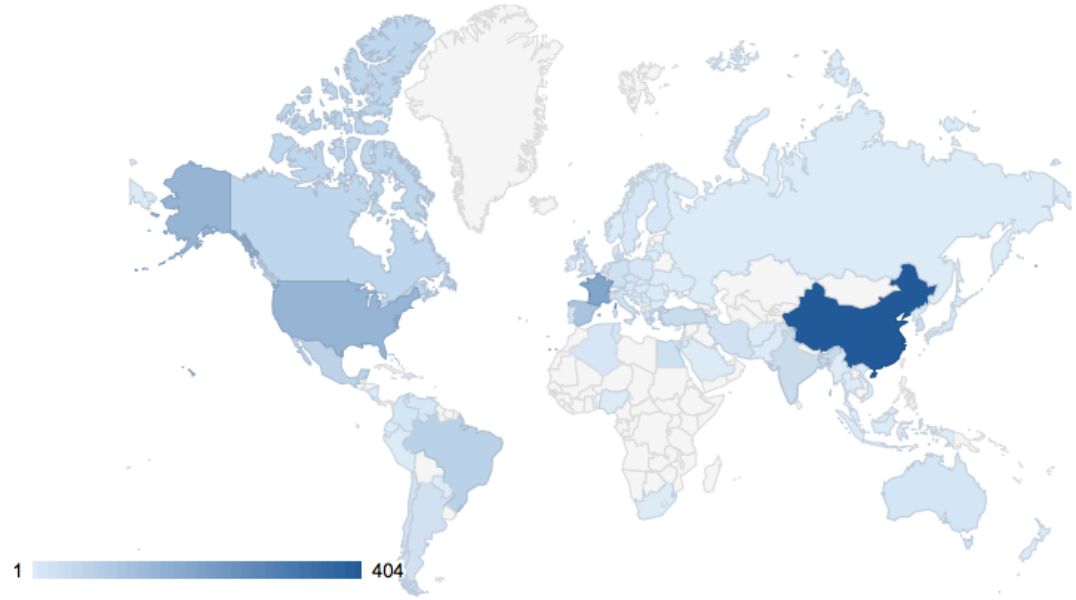

Figure 34:Distribution of volBrain users around the world in January 2018.

3.4 Discussion

In this third chapter, we presented the tools that we developed to preform brain monitoring. First, we detailed the volBrain pipeline mainly based on our PBS framework introduced in Chapter 1. Then, we explained our lifespan model construction based on massive MRI processing. Moreover, we presented our pathological model showing the timeline of brain atrophy in AD. Finally, we described the full open access web-platform that we developed to make our brain monitoring tools worldwide available. Therefore, in this chapter we dealt with two very important aspects of my research – doing translational research by sharing developed tools and producing new knowledge for medicine and neuroscience. We will discuss these both aspect in the following.

In (Manjon and Coupe 2016), we showed that the volBrain pipeline is able to provide *state-of-the-art* results at different levels (intracranial cavity, brain macro-structures and subcortical structures) in a very efficient manner. The proposed pipeline was compared with two well-established software packages (Freesurfer (Fischl 2012) and FIRST (Jenkinson et al. 2012)) for subcortical structure segmentation. The volBrain pipeline was found to significantly improve the accuracy (according to the used protocol) compared to both methods. Regarding to the reproducibility, volBrain was also found to be the more reproducible than Freesurfer and FIRST. This is an important issue since the higher the reproducibility the higher the chances to detect subtle variations induced by the disease. In addition, we found that segmentation masks obtained with FIRST were more accurate and more reproducible than Freesurfer ones. The results on reproducibility between Freesurfer and FIRST were less obvious since they were structure dependent. However, it has to be noted that FIRST failed for 2 cases of 20 (i.e., 10% of failure rate) while both volBrain and Freesurfer worked for all the 20 cases. Finally, the volBrain pipeline is also more computationally efficient than Freesurfer since it takes around 15 minutes to produce the results compared to several hours in the case of Freesurfer (we have to note that Freesurfer provides full brain segmentation and cortical thickness in this time). The volBrain pipeline and FISRT are similar in term of computational time (only for subcortical segmentation without lateral ventricles). We should also remark that volBrain pipeline is one of the few software to provide HIPP segmentation based on EADC protocol which is the new reference for AD. The development of volBrain was the first important step to achieve in order to perform brain monitoring. Afterwards, we started estimating the reference models, the second key element.

In (Coupe et al. 2017), we proposed new reference models for anatomical brain structure based on a massive MRI processing carry out with the volBrain pipeline. This study presented brain volume trajectory over the entire lifespan using the largest age range to date (from few months of life to elderly) and one of the largest number of subjects (N=2944). Beyond proposing new models for brain monitoring, we also produced new knowledges able to help us to answer to one of the main questions in neuroscience – What are normal brain maturation and age-related brain atrophy? Indeed, knowing when brain tissues stop to mature and when they start to degenerate are key questions in neurology (Sowell et al. 2003). In the past, both questions have been treated separately in the literature, preventing us to get a global picture of these join phenomena. In (Coupe et al. 2017), we proposed a unified analysis of brain development and aging at the same time, resulting in new findings as discussed in the following.

First, we showed that the absolute global GM volume followed a complex trajectory with 4 phases: 1) rapid increase from 0 to 8-10y, 2) rapid decrease until 40, 3) a plateau from 40-80y and 4) a rapid decrease after 80y. When global growth effect was corrected, normalized global GM volumes decreased all over lifespan and follow a complex shape with 3 phases: 1) a rapid decrease from 0 to 20y, 2) a plateau from 40 to 80y and 3) a rapid decrease after 80y. This decline of the normalized global GM volume is consistent with the well-known fact that most of the neurogenesis is a prenatal phenomenon (Stiles and Jernigan 2010). In contrast, WM presented a shape close to the usually described inverted U-shape (Walhovd et al. 2011) that persists after controlling for head size. This result indicates that during the early phase of brain development, WM expansion exceeds general growth. The fast simultaneous WM maturation and GM decrease at brain scale from childhood to adolescence are consistent with brain myelination period and cortical thinning process previously observed *ex-vivo* (Huttenlocher and Dabholkar 1997).

Moreover, one of the most marked discrepancy in the literature is about the cortical GM trajectory over childhood (Walhovd et al. 2016). First studies reported an increase with maturation peak in early school age (Giedd et al. 1999, Lenroot et al. 2007, Raznahan et al. 2011). However, mainly monotonic decrease from early childhood have been recently published (Ostby et al. 2009, Brain Development Cooperative 2012, Aubert-Broche et al. 2013, Ducharme et al. 2016, Mills et al. 2016). The first factor that could explain this pronounced divergence is the used volume measurement. In (Coupe et al. 2017), we showed that absolute GM volume follows a 4-stage trajectory with a maturation peak while normalized GM volume follows a 3-stage trajectory exhibiting a decrease all along the lifespan. Therefore, our results are in line with (Giedd et al. 1999, Shaw et al. 2008, Groeschel et al. 2010, Raznahan et al. 2011) using absolute measurement and are consistent with (Ostby et al. 2009, Mills et al. 2016) using normalized measurement. However, several studies reported monotonic decrease using absolute cortical GM volume over childhood (Sowell et al. 2003, Brain Development Cooperative 2012, Aubert-Broche et al. 2013, Ducharme et al. 2016, Mills et al. 2016, Walhovd et al. 2016). This result is in contradiction with studies dedicated to newborn period that report an increase of absolute GM over the first months of life (Groeschel et al. 2010, Holland et al. 2014, Makropoulos et al. 2016). The fact that several studies did not detect GM maturation peak using absolute measurements seems to be related to two main factors, the lack of subjects younger than 5y and the use of low order polynomial models. Indeed, most of the studies presenting monotonic decrease did not include subjects younger than 4y making difficult the detection of GM volume increase over the first years of life. Moreover, this implies that the model fitting was mainly driven by subjects with already mature brains (Sowell et al. 2003, Brain Development Cooperative 2012, Aubert-Broche et al. 2013, Ducharme et al. 2016, Mills et al. 2016, Walhovd et al. 2016). In addition to this potential issue on the used age range, most of these studies were using linear, quadratic or cubic models. Low order polynomial models are not well-designed to capture complex shape such as fast growth with saturation before nonlinear decrease. In (Coupe et al. 2017), we tried to address these two limitations by using subjects younger than 4y old and by considering hybrid models able to handle complex brain change occurring during the first years of life.

Furthermore, deep GM structures are the focus of a great interest due to their important role in various neurodegenerative diseases, and thus have been intensively studied in the past (Fjell et al. 2013). Non-linear trajectories of these structures have been previously described for adulthood (Ziegler et al. 2012, Fjell et al. 2013). More recently, studies taking advantage of the "Big Data sharing" in neuroscience started to analyze subcortical

structure volumes from 20y up to advanced ages to define normative values for adult lifespan (Potvin et al. 2016). However, the limited age range of these studies made impossible to estimate full lifespan models. In (Coupe et al. 2017), we have addressed this important problem by considering subjects covering the entire lifespan. Moreover, we extensively analyzed structure trajectories using both absolute and normalized volumes. Therefore, our results present at the same time the structure maturation peaks occurring during childhood based on absolute volumes and the accelerated atrophy related to aging occurring after 80y obtained using normalized volumes. When deep GM structures are considered at the brain scale, their trajectories present a similar global decrease all along life, except for the medio-temporal regions with a late decrease for amygdala (after 70 years old) and an inverted U-shape for hippocampus. The understanding of the amygdalo-hippocampal complex is important in neurology since it is related to crucial tasks such as memory, spatial navigation or emotional behavior. Moreover, HIPP has been largely studied due to its use as an early biomarker in several neurodegenerative diseases such as AD (Fox et al. 1996, Jack et al. 1997) as already explained in Chapter 2, but also because it is the main location of adult neurogenesis (Eriksson et al. 1998, van Praag et al. 2002). Noteworthy, while amygdala and HIPP are often associated due to their respective contribution to the limbic system, it appears that they present different trajectories. This fact has been previously reported in recent studies (Ziegler et al. 2012, Fjell et al. 2013, Pfefferbaum et al. 2013, Potvin et al. 2016). The long maturation period of the hippocampus may be related to the adult neurogenesis. In fact, it has been shown that neurogenesis in the human hippocampus is substantial until at least the fifth decade of life (Spalding et al. 2013), a finding consistent with our analysis. In contrast to the HIPP, early maturation of the amygdala is consistent with its known function in emotional learning, which allows individuals to avoid aversive events and pursue rewarding experiences (Phelps and LeDoux 2005). The amygdala in humans has been shown to be functional early in life (Tottenham and Sheridan 2009). Our results on amygdala were in accordance with most of the previous studies highlighting a minor effect of aging over adulthood (Walhovd et al. 2011).

In addition to these reference models of normal aging, we presented a study following a similar approach to analyze the timeline of brain atrophy in AD (Coupé et al. 2018). This recent work demonstrated the interest to perform brain monitoring by showing early divergence of pathological and normal model. Indeed, our lifespan analysis based on large-scale datasets using inferred timeline of brain atrophy in AD indicates that the HIPP is the first brain structure (among considered ones) to exhibits a significant volume difference between cognitively normal subjects and subjects who will present clinical symptoms. This difference is detectable early in life, at 39y for the AD/MCI model and at 37y for the AD model. The HIPP is followed by another temporal lobe region, the amygdala, which is different between the two groups at 44y for the AD/MCI model and at 40y for the AD model. It is noticeable that amygdala is undergoing larger changes proportionally to its size compared to HIPP. Finally, the lateral ventricles present an early enlargement at 42y for the AD/MCI model and at 39y for the AD model. However, lateral ventricles enlargement occurring during normal aging reduces the abnormality of this structure after 65y.

Our results presenting the HIPP as the first brain region diverging in the preclinical stage of AD is in accordance with previous morphometric studies focused on the prodromal phase of the disease including ours as discussed in Chapter 2 (den Heijer et al. 2010, Miller et al. 2013, Bernard et al. 2014, Coupé et al. 2015). It is also in accordance with histopathological studies showing the temporal lobe as the starting point of the neurodegenerative process in AD (Braak and Braak 1991). In several long follow-up studies, authors observed that incident cases of AD present

morphometric difference in the hippocampus up to 10 years before the diagnosis (den Heijer et al. 2010, Miller et al. 2013, Bernard et al. 2014, Coupé et al. 2015). In our study, the youngest subjects presenting clinical symptoms included in the AD model are 55 years old, while the pathological trajectories diverge from normal model 18 years before for the AD model and 16 years before for the AD/MCI model. This result suggests that the neurodegeneration of the hippocampus is present several years before the onset of cognitive deficit as discussed in Chapter 2. Therefore, our model seems to confirm the presence of a long-lasting period of silence before the diagnosis of AD, as discussed in (Dubois et al. 2016). Moreover, our model indicates that the age of 40y is a critical period in the onset of the temporal lobe atrophy. Exposure to risk factors (such as diabetes and smoking) occurring at this lifetime period should be considered in future studies to evaluate their implication in the atrophy process (Debette et al. 2011). It worth to note that all the results about HIPP have been obtained using the EADC-ADNI harmonized protocol. Therefore, this study is the largest analysis using this protocol to date.

The second temporal lobe region diverging from the cognitively normal subjects according to our model is the amygdala, which is different from CN at 40y for the AD model and at 44y for the MCI/AD models. Atrophy of this structure has been repeatedly described in AD subjects, with a rate of change less important (Qiu et al. 2009) or similar (Poulin et al. 2011) than hippocampal one. Notably, in the transgenic mouse model of AD, the neurodegeneration in the amygdala even precedes that found in the hippocampus (Lin et al. 2015). In our study, we found that the relative rate of change and abnormality were greater for amygdala than for the HIPP at advanced ages. The early atrophy of the amygdala in the prodromal phase of AD is not surprising when considering the implication of emotion in memory. Indeed, the activity of basolateral and lateral nuclei of the amygdala is associated to a facilitation during the encoding phase and to an enhanced retrieval. These effects being mediated through the important interconnections between these structures and the HIPP (Phelps 2004). In addition, degradation of emotion processing ability is also observed in AD patients, as expected given the amygdala atrophy (Kumfor et al. 2014). Moreover, the atrophy of the amygdala is likely contributing to the olfactory deficits associated with AD, since the cortical nuclei of the amygdala are associated with the processing of olfactory stimuli (De Olmos 2004). Hyposmia has been described in AD (Tabert et al. 2005), and olfactory deficits can substantially precede cognitive symptoms (Djordjevic et al. 2008). However, it has to be taken into account that pathological alterations in AD occur also in other olfactory structures (Ohm and Braak 1987). Finally, timeline atrophy of the HIPP and the amygdala never overlaps across lifespan between the AD and CN models, in contrast to other deep gray matter structures investigated in this study. This result highlights the specificity all along life of the medial temporal lobe atrophy associated to the mnesic symptoms, which characterizes the disease.

According to our results, the volume of the lateral ventricle is also an early biomarker of AD, since its trajectory diverges at 39y for the AD model and at 42y for the AD/MCI model. The potential of lateral ventricle volume as AD biomarkers has been previously mentioned over restricted periods (Nestor et al. 2008). In this study, by analyzing the change of lateral ventricle abnormality over time, we showed that lateral ventricle abnormality decreases after 65y. Therefore, the use of this biomarker is difficult for the late onset cases due to important lateral ventricle enlargement occurring during normal aging. However, it may be useful to discriminate cases around 65y, an early age at which the AD diagnosis is particularly relevant because intervention is more effective in the early phases of the disease (Prince et al. 2011). The importance of taking into account volume increase at advanced age in normal aging has been previously mentioned (Apostolova et al. 2012).

Finally, we decided to share our volBrain pipeline and our reference models by deploying an online web-based platform. The volBrain platform gives access to the whole scientific community not only to our brain monitoring tools but also to our own computational resources. Several platforms are available around the world. Among these platforms we can cite very exciting initiatives such as CBRAIN (Canada), CATI, VIP and FLI (France), NeuGrid (E-U) and SPINE (U.S.A). Each of these platforms follows different philosophy. CBRAIN (Sherif et al. 2015) is not in open access and proposes well-known tools such as FSL, Freesurfer and Minctool packages. VIP (Glatard et al. 2013) is an open access platform also proposing FSL and Freesurfer packages. In the future, pipelines developed by France Life Imaging (FLI)[2] should be integrated within this platform. Another interesting platform is the European platform NeuGrid[3]. NeuGrid is open access and proposes the largest number of tools. However, these tools are mainly based on common tools such as FSL, Freesurfer and SPM packages. In addition, the CATI (Mangin et al. 2014) is the national French platform dedicated to neurodegenerative diseases. CATI is not open access and proposed tools mainly based on Brainvisa[4]. The main advantage of the CATI is to proposed a fully integrated service – from patient recruitment, MRI acquisition to image analysis. Moreover, CATI is built upon a large network of harmonized imaging facilities all around the French territory. Finally, SPINE is an original collaborating platform designed to accelerate scientific discovery while educating non-experts. SPINE is based on crowdsourcing to obtain manual segmentation from non-experts. With volBrain, we developed a different paradigm by proposing an original pipeline in a fully open access as a Software-as-a-Service (SaaS). The volBrain analysis system works remotely through a really simple web interface and automatically produces a report. This report contains reference range to enable brain monitoring, as explained in this chapter. So far, other platforms do not provide such report system for brain monitoring. In 3 years, our volBrain platform processed around 70.000 MRI for 1700 users all around the world. Thanks to our full open access philosophy and a very easy use (one-click solution), our platform grown up rapidly and is now a successful experience[5]. We progressively integrated news tools into our platform. Three years ago, at the beginning, only the volBrain pipeline was available. One year later, we proposed CERES through our platform for cerebellum lobules segmentation. More than 15000 MRI have been processed by the CERES pipeline. Currently, lesionBrain (for lesion extraction) is under beta testing and HIPS (for hippocampus subfields segmentation) is ready for beta testing. Recent funding will help us to deploy a second site at the Bordeaux University. Finally, we are currently moving our architecture to a modular and scalable solution based on Docker technology.

[2] https://www.francelifeimaging.fr
[3] https://neugrid4you.eu
[4] http://brainvisa.info/web/index.html
[5] https://lejournal.cnrs.fr/articles/les-irm-cerebrales-ont-leur-traducteur-automatique

General Discussion

In this habilitation thesis, I presented our work on patch-based MRI analysis dedicated to quantitative MRI, computer-aided diagnosis and brain monitoring. I described the pathway that I followed over the past 10 years from voxelwise analysis to neuroscientific and medical knowledge production. Moreover, I introduced two of our major methodological contributions to medical image analysis – the patch-based segmentation and the patch-based grading. Finally, I detailed our contributions to translational research with the development of the volBrain platform and our last works on anatomical structure trajectory, which provided useful knowledge to better understand brain development and aging, as well as AD progression.

- **Chapter 1**

In Chapter 1, I presented our PBS method based on the nonlocal means strategy. In (Coupe et al. 2010a, Coupe et al. 2011), we were the first to use nonlocal means framework to perform segmentation. Moreover, I showed that our PBS can be used for a large range of applications from brain extraction to lesion detection. At present, PBS is considered as a *state-of-the-art* method and is studied by many groups around the world. Moreover, there exists an international MICCAI workshop dedicated to this topic – the Patch-MI workshop (Wu et al. 2015a, Wu et al. 2016b, Wu et al. 2017) – that I have co-organized in the past 3 years. As explained at the end of Chapter 1, several studies (Wang et al. 2012, Wang et al. 2013a, Wang and Yushkevich 2013), including ours (Manjón et al. 2016, Romero et al. 2016, Romero et al. 2017a, Zandifar et al. 2017), showed that the current optimal PBS strategy is based on a combination of three steps: i) nonlinear multi-atlas warping ii) PBS and iii) error correction involving machine learning methods. This pipeline is competitive for structure segmentation compared to last Deep Learning (DL) strategies as shown in (Wachinger et al. 2017b). In (Wachinger et al. 2017b), the proposed 3D Convolutional Neural Network (CNN) did not outperform a method proposed in 2012 (Wang et al. 2012) based on this framework. Moreover, our CERES2 method was the best ranked method in the international MICCAI challenge ENIGMA, better than several DL methods including Fully Convolution Network (FCN). In addition, our PBS framework remains competitive for MS lesion detection as shown in (Brosch et al. 2016) where FCN obtained worst results than our method (Guizard et al. 2015a). However, patch-wise cascade of CNN (Valverde et al. 2017) demonstrated improved lesion segmentation accuracy compared to our PBS. Patch-wise architecture is well-suited to MS lesion detection since many occurrences of the target object can appear in one training image that can dramatically increase the size of training samples.

According to the recent successes of DL in computer vision, I have no doubt that DL-based methods will soon be a major tool for quantitative MRI analysis – reviews on DL in medical imaging can be found in (Litjens et al. 2017, Shen et al. 2017). However, at present, one of the main issues of using DL in medical imaging is the limited number of training images with manual segmentations. For instance, CNN presented in (Wachinger et al. 2017b) for structure segmentation was tranied on the 30 MRI of the MICCAI challenge 2012 and the CNN-based lesion segmentation methods presented in (Brosch et al. 2016, Valverde et al. 2017) were trained on the 20 MRI of the

MICCAI challenge 2008. While in computer vision, very large labeled databases exist (e.g., ImageNet contains over 10 million of hand-annotated images (Deng et al. 2009)), in medical field the construction of such large dataset is very challenging. Consequently, there is an interesting debate on the choice between exemplar-based methods such as PBS – less demanding in training images – and learning kernel-based methods such as CNN – faster to segment a new case once trained (although it has to be noted that DL training step can require several days on GPU). Currently, I think that the method choice highly depends on the size of the training library. Nevertheless, when the size of training datasets will increase, CNN methods will be more efficient.

Recent ongoing open access initiatives, such as the UKbiobank, have the ambitious goal of imaging more than 100,000 participants (15.000 images are available so far). Moreover, there is a global trend to produce larger datasets. Therefore, we can expect that freely available datasets will be larger in the coming years and thus DL will be a tool of choice. However, the labeling of these large-scale databases will remain an issue. In my opinion, PBS can play an important role to address this problem. As shown in previous works, label propagation (Wolz et al. 2010) based on PBS can produce accurate automatic segmentations over large database using few manually segmented examples. Moreover, our experiments showed that these automatic segmentations can be added to training library to further improve the segmentation accuracy of PBS (Eskildsen et al. 2012, Giraud et al. 2016). Therefore, PBS could be used to label large databases and then DL could be trained on these automatic segmentations. Our first experiments based on this original data augmentation scheme showed promising results using U-Net architecture. Therefore, I am currently working on the combination of PBS and CNN in order to take advantage of both strategies.

- **Chapter 2**

In Chapter 2, I presented our PBG method as an extension of our PBS to perform patient classification. I showed that our PBG can be used for AD diagnosis (Coupe et al. 2012a), AD prognosis at MCI stages (Coupe et al. 2012b, Komlagan et al. 2014, Hett et al. 2016, Hett et al. 2017, Hett et al. 2018) but also at presymptomatic stage (Coupé et al. 2015). Moreover, I described our recent work on multimodal grading of hippocampal subfields as a promising way to better track AD progression (Hett et al. 2018). Finally, the proposed methods comparison showed that our extensions of PBG (Komlagan et al. 2014, Hett et al. 2017) are competitive compared to last DL-based methods (Suk et al. 2017). Indeed, when using similar preprocessing (i.e., nonlinear registration) and ROI (i.e., whole gray matter), PBG and CNN produce similar classification accuracy (Hett et al. 2017). Compared to segmentation, it is easier to get access to larger training datasets in computer-aided diagnosis since manual segmentations are not required. Nevertheless, it seems that the size of currently available datasets (e.g., around 800 subjects for the ADNI1 dataset) is not sufficient to fully take advantage of CNN approaches. Therefore, as mentioned for segmentation, there is an interesting debate to open on the capabilities of exemplar-based methods such as PBG and learning kernel-based methods such as CNN to detect subtle image modifications from "limited" training dataset. In my opinion, with current library size, exemplar-based methods are more sensitive to detect small changes while kernel-based method are more efficient to find general class-specific features. Consequently, further research would be very interesting on this aspect, in particular on the combination of the two approaches to take advantage of the sensitivity of PBG and the specificity of CNN.

As showed in Chapter 3, one way to address the problem of the training dataset size in computer-aided diagnosis is to combine several datasets. To make the aggregation strategy of these datasets effective, we will need to homogenize images quality and contrast. Therefore, preprocessing steps – such as denoising, intensity normalization and so on – will be crucial to successfully achieve such aggregation. As discussed in (Akkus et al. 2017), despite their robustness, an efficient preprocessing pipeline remains very important when using DL methods in medical imaging. Thanks to our works dedicated to the volBrain platform, we have acquired a great expertise in the preprocessing of multiple large datasets. Moreover, as presented in Chapter 3, we also learned to aggregate multiple heterogeneous datasets when we estimated our lifespan model. Therefore, I think that our expertise on low level task such as denoising, registration and intensity normalization will be really useful to build large-scale databases and thus to address new challenges related to DL and Big Data.

- **Chapter 3**

In Chapter 3, I presented our works on brain monitoring, or at least on our efforts to make it possible one day. In this chapter, I first described the volBrain pipeline. I showed that our software is very competitive for subcortical structure segmentation in terms of accuracy, reproducibility and robustness compared to *state-of-the-art* software such as Freesurfer (developed by Harvard and the M.I.T) and FIRST (developed by the University of Oxford). Moreover, I presented our lifespan models that revealed for the first time the brain trajectory from few months of life to advanced ages. In addition, I detailed our pathological model enabling to tack brain alterations caused by AD across the entire lifespan. Finally, I explained the fully open access philosophy of our volBrain platform and its translational impact. After 3 years of experiences, the volBrain platform already processed more than 70.000 MRI for more than 800 institutions around the world. This usage statistic demonstrates the need of proposing online service dedicated to automatic brain analysis and brain monitoring. In addition, this great success prompts us to rethink its architecture by using scalable technology such as containerization (e.g., Docker).

With the "volBrain adventure" and the emergence of very large dataset freely available, I have faced several challenges related to Big Data as defined by the usual 3Vs model (Volume, Variety and Velocity). My first challenge was to deal with large volumes of data and thus to manage their storage, their processing, their traceability, etc. To address this challenge, we proposed the fully automatic volBrain platform based on local servers hosted at the Universitat Politecnica de Valencia. At present, we are currently deploying a second site at the University of Bordeaux to increase our computational and storage capabilities. This will temporarily solve the issue of dealing with large volumes of data. However, in the future, cloud-based strategy seems a promising solution to enable scalable data storage and processing. The migration of volBrain to cloud-based solution will be facilitated by the new architecture of volBrain.

My second challenge was to propose a fast and robust pipeline to make the processing of large databases practical and efficient. With our PBS framework based on OPAL, we drastically decreased multi-atlas segmentation from several hours to few minutes. This time can be potential reduced in the near future thanks to GPU-based methods such as DL. In addition to velocity, robustness is also important when processing large databases. Thanks to its simplicity, PBS demonstrated a good robustness with less than 2% of failure through the web platform. Despite

this low failure rate, one aspect that I had neglected at the beginning was the quality control (QC). When I worked on lifespan models, I spent a significant number of evenings to manually perform human-based QC. In my opinion, automatic QC of image processing pipeline is clearly under considered. Until recently, most of publications dedicated to brain MRI study involved several tens or hundreds of subjects and thus manual QC could be performed. However, when we will deal with hundreds of thousands – already 70.000 MRI processed by volBrain and soon 100.000 subjects available with the UKbiobank – QC based on visual inspection will be no more realistic. Therefore, I think that one part of my future investigations will be dedicated to automatic QC.

The last V of the usual Big Data 3Vs model is the variety. So far, the data variety managed by the volBrain platform is limited. The user can optionally provide the subject's age and gender in addition to a T1w MRI. Thanks to the integration of lesionBrain and HIPS pipelines, several modalities of MRI are also managed. However, in my opinion, this variety will still not be sufficient to take full advantage of Big Data promises. I think that in medical imaging one emerging and promising field is the imaging genetics (Medland et al. 2014). Imaging genetics is a recent field of research that aims to identify correlations between genetic variants and anatomical or functional brain characteristics. Therefore, imaging genetics enables to find relationships between a genotype and an imaging phenotype. In my opinion, this field has a great potential but requires to process large volumes of highly heterogeneous data (i.e., imaging and genetic) in a practical time. My long-term project would be to combine the volBrain platform with an existing genetic platform to perform such imaging genetics.

Recently, two Vs have been added to the usual 3Vs Big Data model – Veracity and Value – to define the new 5Vs Big Data model. In medical imaging, the reliability of the used data is related to the questions of QC and traceability. Unlike metadata provided by freely available datasets, where veracity is usually guaranteed, metadata provided by users to the volBrain platform are not trusty. This is one important limitation of the fully open access web-based approach. One way to answer this question is to process freely available datasets by ourselves through the web-platform as we did in (Coupe et al. 2017, Coupé et al. 2018). Another way to address this issue is to directly connect the platform to a PACS. However, such strategy implies that the platform is no more open access and raises several security issues. The current available platforms (e.g., CATI, FLI, VIP, CBRAIN, SPINE, etc.) follow different ways to address the veracity problem such as using private internal database (e.g., CATI) or crowdsourcing (e.g., SPINE). Therefore, it will be interesting to analyze feedback from each of them and to discuss possible connections between these platforms. The last V is the value of these volumes of data. In (Coupe et al. 2017, Coupé et al. 2018), we already demonstrated the high value of using large dataset. Moreover, as previously mentioned, large dataset will also be very useful for DL methods. It is obvious that the emergence of very large multimodal datasets (e.g., UKbiobank) is an invaluable opportunity for medical and neuroscientific researches. However, this next generation of dataset will also require tools designed to manage them such as the volBrain platform.

To conclude this manuscript describing my last 8 years devoted to medical image analysis, I would like to say that it was a real pleasure to work in this multidisciplinary field. I had the chance to meet passionate experts from different domains ranging from mathematics to medicine who gave me a little of their time to explain to me the specific concepts that I needed for my research. I hope to continue this adventure as long as possible and I look forward to the next challenges related to A.I. and BigData in neuroimaging.

References

3C-Study-Group (2003). "Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population." Neuroepidemiology **22**(6): 316-325.

Ahmed, M. N., S. M. Yamany, N. Mohamed, A. A. Farag and T. Moriarty (2002). "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data." IEEE transactions on medical imaging **21**(3): 193-199.

Aizenstein, H. J., R. D. Nebes, J. A. Saxton, J. C. Price, C. A. Mathis, N. D. Tsopelas, S. K. Ziolko, J. A. James, B. E. Snitz, P. R. Houck, W. Bi, A. D. Cohen, B. J. Lopresti, S. T. DeKosky, E. M. Halligan and W. E. Klunk (2008). "Frequent amyloid deposition without significant cognitive impairment among the elderly." Archives of Neurology **65**(11): 1509-1517.

Akkus, Z., A. Galimzianova, A. Hoogi, D. L. Rubin and B. J. Erickson (2017). "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions." Journal of Digital Imaging: 1-11.

Aljabar, P., R. A. Heckemann, A. Hammers, J. V. Hajnal and D. Rueckert (2009). "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy." NeuroImage **46**(3): 726-738.

Apostolova, L. G., R. A. Dutton, I. D. Dinov, K. M. Hayashi, A. W. Toga, J. L. Cummings and P. M. Thompson (2006). "Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps." Archives of neurology **63**(5): 693-699.

Apostolova, L. G., A. E. Green, S. Babakchanian, K. S. Hwang, Y.-Y. Chou, A. W. Toga and P. M. Thompson (2012). "Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and Alzheimer's disease." Alzheimer disease and associated disorders **26**(1): 17.

Ashburner, J. (2012). "SPM: a history." Neuroimage **62**(2): 791-800.

Ashburner, J. and K. J. Friston (2000). "Voxel-based morphometry--the methods." Neuroimage **11**(6 Pt 1): 805-821.

Asman, A. J. and B. A. Landman (2013). "Non-local statistical label fusion for multi-atlas segmentation." Medical image analysis **17**(2): 194-208.

Aubert-Broche, B., V. S. Fonov, D. Garcia-Lorenzo, A. Mouiha, N. Guizard, P. Coupe, S. F. Eskildsen and D. L. Collins (2013). "A new method for structural volume analysis of longitudinal brain MRI data and its application in studying the growth trajectories of anatomical brain structures in childhood." Neuroimage **82**: 393-402.

Avants, B. B., N. J. Tustison, G. Song, P. A. Cook, A. Klein and J. C. Gee (2011). "A reproducible evaluation of ANTs similarity metric performance in brain image registration." Neuroimage **54**(3): 2033-2044.

Bai, W., W. Shi, C. Ledig and D. Rueckert (2015). "Multi-atlas segmentation with augmented features for cardiac MR images." Medical image analysis **19**(1): 98-109.

Bai, W., W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters and D. Rueckert (2013). "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images." IEEE transactions on medical imaging **32**(7): 1302-1315.

Ball, M. J., M. Fisman, V. Hachinski, W. Blume, A. Fox, V. A. Kral, A. J. Kirshen, H. Fox and H. Merskey (1985). "A new definition of Alzheimer's disease: a hippocampal dementia." Lancet **1**(8419): 14-16.

Ballard, C., Z. Khan, H. Clack and A. Corbett (2011). "Nonpharmacological treatment of Alzheimer disease." Canadian journal of psychiatry. Revue canadienne de psychiatrie **56**(10): 589-595.

Barnes, C., E. Shechtman, A. Finkelstein and D. Goldman (2009). "PatchMatch: A randomized correspondence algorithm for structural image editing." ACM Transactions on Graphics-TOG **28**(3): 24.

Barnes, C., E. Shechtman, D. Goldman and A. Finkelstein (2010). "The generalized patchmatch correspondence algorithm." Computer Vision–ECCV 2010: 29-43.

Barnes, J., J. Foster, R. G. Boyes, T. Pepple, E. K. Moore, J. M. Schott, C. Frost, R. I. Scahill and N. C. Fox (2008). "A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus." NeuroImage **40**(4): 1655-1671.

Bernard, C., C. Helmer, B. Dilharreguy, H. Amieva, S. Auriacombe, J.-F. Dartigues, M. Allard and G. Catheline (2014). "Time course of brain volume changes in the preclinical phase of Alzheimer's disease." Alzheimer's & Dementia **10**(2): 143-151.

Bernasconi, N., A. Bernasconi, Z. Caramanos, S. B. Antel, F. Andermann and D. L. Arnold (2003). "Mesial temporal damage in temporal lobe epilepsy: a volumetric MRI study of the hippocampus, amygdala and parahippocampal region." Brain **126**(Pt 2): 462-469.

Braak, E. and H. Braak (1997). "Alzheimer's disease: transiently developing dendritic changes in pyramidal cells of sector CA1 of the Ammon's horn." Acta neuropathologica **93**(4): 323-325.

Braak, H., I. Alafuzoff, T. Arzberger, H. Kretzschmar and K. Del Tredici (2006). "Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry." Acta neuropathologica **112**(4): 389-404.

Braak, H. and E. Braak (1991). "Neuropathological stageing of Alzheimer-related changes." Acta neuropathologica **82**(4): 239-259.

Brain Development Cooperative, G. (2012). "Total and regional brain volumes in a population-based normative sample from 4 to 18 years: the NIH MRI Study of Normal Brain Development." Cereb Cortex **22**(1): 1-12.

Braskie, M. N. and P. M. Thompson (2014). "A focus on structural brain imaging in the Alzheimer's disease neuroimaging initiative." Biol Psychiatry **75**(7): 527-533.

Bremner, J. D., M. Narayan, E. R. Anderson, L. H. Staib, H. L. Miller and D. S. Charney (2000). "Hippocampal volume reduction in major depression." Am J Psychiatry **157**(1): 115-118.

Bremner, J. D., P. Randall, T. M. Scott, R. A. Bronen, J. P. Seibyl, S. M. Southwick, R. C. Delaney, G. McCarthy, D. S. Charney and R. B. Innis (1995). "MRI-based measurement of hippocampal volume in patients with combat-related posttraumatic stress disorder." Am J Psychiatry **152**(7): 973-981.

Bron, E. E., M. Smits, W. M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. Steketee, C. Mendez Orellana, R. Meijboom, M. Pinto, J. R. Meireles, C. Garrett, A. J. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cardenas-Pena, A. M. Alvarez-Meza, C. V. Dolph, K. M. Iftekharuddin, S. F. Eskildsen, P. Coupe, V. S. Fonov, K. Franke, C. Gaser, C. Ledig, R. Guerrero, T. Tong, K. R. Gray, E. Moradi, J. Tohka, A. Routier, S. Durrleman, A. Sarica, G. Di Fatta, F. Sensi, A. Chincarini, G. M. Smith, Z. V. Stoyanov, L. Sorensen, M. Nielsen, S. Tangaro, P. Inglese, C. Wachinger, M. Reuter, J. C. van Swieten, W. J. Niessen, S. Klein and I. Alzheimer's Disease Neuroimaging (2015). "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge." Neuroimage **111**: 562-579.

Brosch, T., L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee and R. Tam (2016). "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation." IEEE transactions on medical imaging **35**(5): 1229-1239.

Buades, A., B. Coll and J. M. Morel (2005). "A non-local algorithm for image denoising." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 2, Proceedings: 60-65.

Buades, A., B. Coll and J. M. Morel (2010). "Image Denoising Methods. A New Nonlocal Principle." Siam Review **52**(1): 113-147.

Buss, C., C. Lord, M. Wadiwalla, D. H. Hellhammer, S. J. Lupien, M. J. Meaney and J. C. Pruessner (2007). "Maternal care modulates the relationship between prenatal risk and hippocampal volume in women but not in men." J Neurosci **27**(10): 2592-2595.

Caldairou, B., B. C. Bernhardt, J. Kulaga-Yoskovitz, H. Kim, N. Bernasconi and A. Bernasconi (2016). "A Surface Patch-Based Segmentation Method for Hippocampal Subfields." International Conference on Medical Image Computing and Computer-Assisted Intervention: 379-387.

Callaway, E. (2012). "Alzheimer's drugs take a new tack." Nature **489**: 13-14.

Carass, A., J. L. Cuzzocreo, S. Han, C. R. Hernandez-Castillo, P. E. Rasser, M. Ganz, V. Beliveau, M. Price, V. A. Cardenas, G. Fein, J. Dolz, I. Ben Ayed, C. Desrosiers, B. Thyreau, J. E. Romero, P. Coupé, J. V. Manjón, V. S. Fonov, D. L. Collins, S. H. Ying, D. Crocetti, B. A. Landman, S. H. M. ostofsky, P. A. T. hompson and J. L. Prince (2018). "Comparing fully automated state-of-the-art cerebellum parcellation from Magnetic Resonance Imaging." Sbmitted to Neuroimage.

Carlesimo, G. A., F. Piras, M. D. Orfei, M. Iorio, C. Caltagirone and G. Spalletta (2015). "Atrophy of presubiculum and subiculum is the earliest hippocampal anatomical marker of Alzheimer's disease." Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **1**(1): 24-32.

Chakravarty, M. M., P. Steadman, M. C. Eede, R. D. Calcott, V. Gu, P. Shaw, A. Raznahan, D. L. Collins and J. P. Lerch (2013). "Performing label-fusion-based segmentation using multiple automatically generated templates." Human brain mapping **34**(10): 2635-2654.

Chetelat, G., V. L. Villemagne, P. Bourgeat, K. E. Pike, G. Jones, D. Ames, K. A. Ellis, C. Szoeke, R. N. Martins, G. J. O'Keefe, O. Salvado, C. L. Masters and C. C. Rowe (2010). "Relationship between atrophy and beta-amyloid deposition in Alzheimer disease." Annals of Neurology **67**(3): 317-324.

Chupin, M., E. Gerardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehericy, H. Benali, L. Garnero and O. Colliot (2009). "Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI." Hippocampus **19**(6): 579-587.

Chupin, M., A. R. Mukuna-Bantumbakulu, D. Hasboun, E. Bardinet, S. Baillet, S. Kinkingnehun, L. Lemieux, B. Dubois and L. Garnero (2007). "Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with Alzheimer's disease." NeuroImage **34**(3): 996-1019.

Clerx, L., I. A. van Rossum, L. Burns, D. L. Knol, P. Scheltens, F. Verhey, P. Aalten, P. Lapuerta, L. van de Pol, R. van Schijndel, R. de Jong, F. Barkhof, R. Wolz, D. Rueckert, M. Bocchetta, M. Tsolaki, F. Nobili, L. O. Wahlund, L. Minthon, L. Frolich, H. Hampel, H. Soininen and P. J. Visser (2013). "Measurements of medial temporal lobe atrophy for prediction of Alzheimer's disease in subjects with mild cognitive impairment." Neurobiol Aging **34**(8): 2003-2013.

Collins, D. L., C. J. Holmes, T. M. Peters and A. C. Evans (1995). "Automatic 3-D model-based neuroanatomical segmentation." Human Brain Mapping **3**(3): 190-208.

Collins, D. L. and J. C. Pruessner (2010). "Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion." Neuroimage **52**(4): 1355-1366.

Cordier, N., H. Delingette and N. Ayache (2016a). "A patch-based approach for the segmentation of pathologies: Application to glioma labelling." IEEE transactions on medical imaging **35**(4): 1066-1076

Cordier, N., H. Delingette, M. Lê and N. Ayache (2016b). "Extended modality propagation: image synthesis of pathological cases." IEEE transactions on medical imaging **35**(12): 2598-2608.

Coupe, P., G. Catheline, E. Lanuza, J. V. Manjon and I. Alzheimer's Disease Neuroimaging (2017). "Towards a unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis." Hum Brain Mapp **38**(11): 5501-5518.

Coupe, P., S. F. Eskildsen, J. V. Manjon, V. S. Fonov, D. L. Collins and I. Alzheimer's disease Neuroimaging (2012a). "Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease." Neuroimage **59**(4): 3736-3747.

Coupe, P., S. F. Eskildsen, J. V. Manjon, V. S. Fonov, J. C. Pruessner, M. Allard, D. L. Collins and I. Alzheimer's Disease Neuroimaging (2012b). "Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease." Neuroimage Clin **1**(1): 141-152.

Coupé, P., V. S. Fonov, C. Bernard, A. Zandifar, S. F. Eskildsen, C. Helmer, J. V. Manjón, H. Amieva, J. F. Dartigues and M. Allard (2015). "Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: Toward an early individual prognosis." Human brain mapping **36**(12): 4758-4770.

Coupe, P., P. Hellier, C. Kervrann and C. Barillot (2009a). "Nonlocal means-based speckle filtering for ultrasound images." IEEE Trans Image Process **18**(10): 2221-2229.

Coupe, P., P. Hellier, S. Prima, C. Kervrann and C. Barillot (2008a). "3D wavelet subbands mixing for image denoising." Int J Biomed Imaging **2008**: 590183.

Coupe, P., J. V. Manjon, M. Chamberland, M. Descoteaux and B. Hiba (2013). "Collaborative patch-based super-resolution for diffusion-weighted images." NeuroImage **83**: 245-261.

Coupe, P., J. V. Manjon, V. Fonov, J. Pruessner, M. Robles and D. L. Collins (2010a). "Nonlocal patch-based label fusion for hippocampus segmentation." Med Image Comput Comput Assist Interv **13**(Pt 3): 129-136.

Coupe, P., J. V. Manjon, V. Fonov, J. Pruessner, M. Robles and D. L. Collins (2011). "Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation." Neuroimage **54**(2): 940-954.

Coupe, P., J. V. Manjon, E. Gedamu, D. Arnold, M. Robles and D. L. Collins (2009b). "An object-based method for Rician noise estimation in MR images." Med Image Comput Comput Assist Interv **12**(Pt 2): 601-608.

Coupe, P., J. V. Manjon, E. Gedamu, D. Arnold, M. Robles and D. L. Collins (2010b). "Robust Rician noise estimation for MR images." Med Image Anal **14**(4): 483-493.

Coupé, P., J. V. Manjón, E. Lanuza and G. Catheline (2018). "TIMELINE OF BRAIN ALTERATIONS IN ALZHEIMER'S DISEASE ACROSS THE ENTIRE LIFESPAN." PNAS: under review.

Coupe, P., M. Munz, J. V. Manjon, E. S. Ruthazer and D. L. Collins (2012c). "A CANDLE for a deeper in vivo insight." Med Image Anal **16**(4): 849-864.

Coupe, P., P. Yger and C. Barillot (2006). "Fast non local means denoising for 3D MR images." Med Image Comput Comput Assist Interv **9**(Pt 2): 33-40.

Coupe, P., P. Yger, S. Prima, P. Hellier, C. Kervrann and C. Barillot (2008b). "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images." IEEE transactions on medical imaging **27**(4): 425-441.

Criminisi, A., P. Perez and K. Toyama (2004). "Region filling and object removal by exemplar-based image inpainting." IEEE Trans Image Process **13**(9): 1200-1212.

Cuingnet, R., E. Gerardin, J. Tessieras, G. Auzias, S. Lehericy, M. O. Habert, M. Chupin, H. Benali, O. Colliot and I. Alzheimer's Disease Neuroimaging (2011). "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database." Neuroimage **56**(2): 766-781.

Cummings, J. L., G. T. Grossberg, P. Green, M. Tocco, E. M. Resnick, J. T. Olin and M. Graham (2007). "A review of the safety and tolerability of treatments for moderate to severe Alzheimer's disease." European journal of neurology **14**: 52-52.

Davatzikos, C., P. Bhatt, L. M. Shaw, K. N. Batmanghelich and J. Q. Trojanowski (2011). "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification." Neurobiol Aging **32**(12): 2322 e2319-2327.

de Jong, L. W., K. van der Hiele, I. M. Veer, J. J. Houwing, R. G. Westendorp, E. L. Bollen, P. W. de Bruin, H. A. Middelkoop, M. A. van Buchem and J. van der Grond (2008). "Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study." Brain : a journal of neurology **131**(Pt 12): 3277-3285.

De Leon, M., A. George, L. Stylopoulos, G. Smith and D. Miller (1989). "EARLY MARKER FOR ALZHEIMER'S DISEASE: THE ATROPHIC HIPPOCAMPUS." The Lancet **334**(8664): 672-673.

De Olmos, J. S. (2004). Amygdala. The Human Nervous System (Second Edition)**:** 739-868.

Debette, S., S. Seshadri, A. Beiser, R. Au, J. Himali, C. Palumbo, P. Wolf and C. DeCarli (2011). "Midlife vascular risk factor exposure accelerates structural brain aging and cognitive decline." Neurology **77**(5): 461-468.

den Heijer, T., F. van der Lijn, P. J. Koudstaal, A. Hofman, A. van der Lugt, G. P. Krestin, W. J. Niessen and M. M. Breteler (2010). "A 10-year follow-up of hippocampal volume on magnetic resonance imaging in early dementia and cognitive decline." Brain **133**(4): 1163-1172.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on: 248-255.

Devanand, D. P., R. Bansal, J. Liu, X. Hao, G. Pradhaban and B. S. Peterson (2012). "MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease." NeuroImage **60**(3): 1622-1629.

Diedrichsen, J. (2006). "A spatially unbiased atlas template of the human cerebellum." Neuroimage **33**(1): 127-138.

Djordjevic, J., M. Jones-Gotman, K. De Sousa and H. Chertkow (2008). "Olfaction in patients with mild cognitive impairment and Alzheimer's disease." Neurobiology of aging **29**(5): 693-706.

Dolz, J., C. Desrosiers and I. B. Ayed (2018). "Comparing fully automated state-of-the-art cerebellum parcellation from Magnetic Resonance Imaging: Cerebellum parcellation from a deep learning perspective." Submitted to NeuroImage.

Driscoll, I., D. A. Hamilton, H. Petropoulos, R. A. Yeo, W. M. Brooks, R. N. Baumgartner and R. J. Sutherland (2003). "The aging hippocampus: cognitive, biochemical and structural findings." Cerebral cortex **13**(12): 1344-1351.

Du, A. T., N. Schuff, D. Amend, M. P. Laakso, Y. Y. Hsu, W. J. Jagust, K. Yaffe, J. H. Kramer, B. Reed, D. Norman, H. C. Chui and M. W. Weiner (2001). "Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease." J Neurol Neurosurg Psychiatry **71**(4): 441-447.

Dubois, B., H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram and K. Blennow (2016). "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria." Alzheimer's & Dementia **12**(3): 292-323.

Ducharme, S., M. D. Albaugh, T. V. Nguyen, J. J. Hudziak, J. M. Mateos-Perez, A. Labbe, A. C. Evans, S. Karama and G. Brain Development Cooperative (2016). "Trajectories of cortical thickness maturation in normal brain development--The importance of quality control procedures." Neuroimage **125**: 267-279.

Duchesne, S., J. Pruessner and D. L. Collins (2002). "Appearance-based segmentation of medial temporal lobe structures." Neuroimage **17**(2): 515-531.

Dyrba, M., F. Barkhof, A. Fellgiebel, M. Filippi, L. Hausner, K. Hauenstein, T. Kirste and S. J. Teipel (2015). "Predicting Prodromal Alzheimer's Disease in Subjects with Mild Cognitive Impairment Using Machine Learning Classification of Multimodal Multicenter Diffusion-Tensor and Magnetic Resonance Imaging Data." Journal of Neuroimaging **25**(5): 738-747.

Efros, A. A. and W. T. Freeman (2001). Image quilting for texture synthesis and transfer. Proceedings of the 28th annual conference on Computer graphics and interactive techniques, ACM**:** 341-346.

Eriksson, P. S., E. Perfilieva, T. Björk-Eriksson, A.-M. Alborn, C. Nordborg, D. A. Peterson and F. H. Gage (1998). "Neurogenesis in the adult human hippocampus." Nature medicine **4**(11): 1313-1317.

Eskildsen, S. F., P. Coupe, V. Fonov, J. V. Manjon, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Ostergaard, D. L. Collins and I. Alzheimer's Disease Neuroimaging (2012). "BEaST: brain extraction based on nonlocal segmentation technique." Neuroimage **59**(3): 2362-2373.

Eskildsen, S. F., P. Coupe, D. Garcia-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins and I. Alzheimer's Disease Neuroimaging (2013). "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning." Neuroimage **65**: 511-521.

Eskildsen, S. F., P. Coupe, D. Garcia-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins and I. Alzheimer's Disease Neuroimaging (2015). "Structural imaging biomarkers of Alzheimer's disease: predicting disease progression." Neurobiology of aging **36**: S23-S31.

Evans, A. C. and B. D. C. Group (2006). "The NIH MRI study of normal brain development." Neuroimage **30**(1): 184-202.

Ewers, M., C. Walsh, J. Q. Trojanowski, L. M. Shaw, R. C. Petersen, C. R. Jack, Jr., H. H. Feldman, A. L. Bokde, G. E. Alexander, P. Scheltens, B. Vellas, B. Dubois, M. Weiner and H. Hampel (2012). "Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance." Neurobiology of Aging **33**(7): 1203-1214.

Fellgiebel, A., P. R. Dellani, D. Greverus, A. Scheurich, P. Stoeter and M. J. Müller (2006). "Predicting conversion to dementia in mild cognitive impairment by volumetric and diffusivity measurements of the hippocampus." Psychiatry Research: Neuroimaging **146**(3): 283-287.

Fellgiebel, A. and I. Yakushev (2011). "Diffusion tensor imaging of the hippocampus in MCI and early Alzheimer's disease." Journal of Alzheimer's Disease **26**(s3): 257-262.

Fischl, B. (2012). "FreeSurfer." Neuroimage **62**(2): 774-781.

Fischl, B., D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen and A. M. Dale (2002). "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain." Neuron **33**(3): 341-355.

Fjell, A. M., L. T. Westlye, H. Grydeland, I. Amlien, T. Espeseth, I. Reinvang, N. Raz, D. Holland, A. M. Dale, K. B. Walhovd and I. Alzheimer Disease Neuroimaging (2013). "Critical ages in the life course of the adult brain: nonlinear subcortical aging." Neurobiol Aging **34**(10): 2239-2247.

Folstein, M. F., S. E. Folstein and P. R. McHugh (1975). ""Mini-mental state". A practical method for grading the cognitive state of patients for the clinician." Journal of psychiatric research **12**(3): 189-198.

Fonov, V., P. Coupé, S. Eskildsen, J. Manjon and L. Collins (2012). Multi-atlas labeling with population-specific template and non-local patch-based label fusion. MICCAI 2012 Workshop on multi-atlas labeling.

Fonov, V. S. and D. L. Collins (2018). "Comparing fully automated state-of-the-art cerebellum parcellation from Magnetic Resonance Imaging: U-Net Parcellation of the Cerebellum." Submitted to NeuroImage.

Fox, N., E. Warrington, P. Freeborough, P. Hartikainen, A. Kennedy, J. Stevens and M. N. Rossor (1996). "Presymptomatic hippocampal atrophy in Alzheimer's disease." Brain **119**(6): 2001-2007.

Franke, K., E. Luders, A. May, M. Wilke and C. Gaser (2012). "Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI." Neuroimage **63**(3): 1305-1312.

Freund, Y. and R. E. Schapire (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. European conference on computational learning theory, Springer.

Frisoni, G. B., N. C. Fox, C. R. Jack, P. Scheltens and P. M. Thompson (2010). "The clinical use of structural MRI in Alzheimer disease." Nature Reviews Neurology **6**(2): 67-77.

Frisoni, G. B., C. R. Jack, M. Bocchetta, C. Bauer, K. S. Frederiksen, Y. Liu, G. Preboske, T. Swihart, M. Blair and E. Cavedo (2015). "The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity." Alzheimer's & Dementia **11**(2): 111-125.

Gainotti, G., A. Acciarri, A. Bizzarro, C. Marra, C. Masullo, S. Misciagna, T. Tartaglione, A. Valenza and C. Colosimo (2004). "The role of brain infarcts and hippocampal atrophy in subcortical ischaemic vascular dementia." Neurol Sci **25**(4): 192-197.

Geremia, E., O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi and N. Ayache (2011). "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images." NeuroImage **57**(2): 378-390.

Ghafoorian, M., N. Karssemeijer, T. Heskes, I. W. M. Uden, C. I. Sanchez, G. Litjens, F.-E. Leeuw, B. Ginneken, E. Marchiori and B. Platel (2017). "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities." Scientific Reports **7**(1): 5110.

Ghanei, A., H. Soltanian-Zadeh and J. P. Windham (1998). "Segmentation of the hippocampus from brain MRI using deformable contours." Comput Med Imaging Graph **22**(3): 203-216.

Giedd, J. N., J. Blumenthal, N. O. Jeffries, F. X. Castellanos, H. Liu, A. Zijdenbos, T. Paus, A. C. Evans and J. L. Rapoport (1999). "Brain development during childhood and adolescence: a longitudinal MRI study." Nat Neurosci **2**(10): 861-863.

Giedd, J. N. and J. L. Rapoport (2010). "Structural MRI of pediatric brain development: what have we learned and where are we going?" Neuron **67**(5): 728-734.

Giraud, R., V. T. Ta, N. Papadakis, J. V. Manjon, D. L. Collins, P. Coupe and I. Alzheimer's Disease Neuroimaging (2016). "An Optimized PatchMatch for multi-scale and multi-feature label fusion." Neuroimage **124**(Pt A): 770-782.

Glatard, T., C. Lartizien, B. Gibaud, R. Ferreira da Silva, G. Forestier, F. Cervenansky, M. Alessandrini, H. Benoit-Cattin, O. Bernard and S. Camarasu-Pop (2013). "A virtual imaging platform for multi-modality medical image simulation." Medical Imaging, IEEE Transactions on **32**(1): 110-118.

Good, C. D., I. S. Johnsrude, J. Ashburner, R. N. Henson, K. Fristen and R. S. Frackowiak (2002). A voxel-based morphometric study of ageing in 465 normal adult human brains. Biomedical Imaging, 2002. 5th IEEE EMBS International Summer School on, IEEE.

Gousias, I. S., D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards and A. Hammers (2008). "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest." NeuroImage **40**(2): 672-684.

Groeschel, S., B. Vollmer, M. King and A. Connelly (2010). "Developmental changes in cerebral grey and white matter volume from infancy to adulthood." International Journal of Developmental Neuroscience **28**(6): 481-489.

Guerrero, R., C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. C. Valdés-Hernández, D. A. Dickie and J. Wardlaw (2018). "White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks." NeuroImage: Clinical **17**: 918-934.

Guizard, N., P. Coupe, V. S. Fonov, J. V. Manjon, D. L. Arnold and D. L. Collins (2015a). "Rotation-invariant multi-contrast non-local means for MS lesion segmentation." Neuroimage Clin **8**: 376-389.

Guizard, N., K. Nakamura, P. Coupe, V. S. Fonov, D. L. Arnold and D. L. Collins (2015b). "Non-Local Means Inpainting of MS Lesions in Longitudinal Image Processing." Front Neurosci **9**: 456.

Haegelen, C., P. Coupe, V. Fonov, N. Guizard, P. Jannin, X. Morandi and D. L. Collins (2013). "Automated segmentation of basal ganglia and deep brain structures in MRI of Parkinson's disease." Int J Comput Assist Radiol Surg **8**(1): 99-110.

Hammers, A., R. Allom, M. J. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks and J. S. Duncan (2003). "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe." Human brain mapping **19**(4): 224-247.

Hammers, A., R. Heckemann, M. J. Koepp, J. S. Duncan, J. V. Hajnal, D. Rueckert and P. Aljabar (2007). "Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: a proof-of-principle study." NeuroImage **36**(1): 38-47.

Hammers, A., M. J. Koepp, S. L. Free, M. Brett, M. P. Richardson, C. Labbé, V. J. Cunningham, D. J. Brooks and J. Duncan (2002). "Implementation and application of a brain template for multiple volumes of interest." Human brain mapping **15**(3): 165-174.

Heckemann, R. A., J. V. Hajnal, P. Aljabar, D. Rueckert and A. Hammers (2006). "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion." NeuroImage **33**(1): 115-126.

Heckemann, R. A., S. Keihaninejad, P. Aljabar, D. Rueckert, J. V. Hajnal, A. Hammers and I. Alzheimer's Disease Neuroimaging (2010). "Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation." Neuroimage **51**(1): 221-227.

Hett, K., V.-T. Ta, G. l. Catheline, T. Tourdias, J. V. Manjón and P. Coupé (2018). "Multimodal Hippocampal Subfields Biomarkers For Alzheimer's Disease Tracking." Submitted to Scientific Report.

Hett, K., V.-T. Ta, R. Giraud, M. Mondino, J. V. Manjón and P. Coupé (2016). "Patch-based DTI grading: Application to Alzheimer's disease classification." International Workshop on Patch-based Techniques in Medical Imaging: 76-83.

Hett, K., V.-T. Ta, J. V. Manjón and P. Coupé (2017). "Adaptive Fusion of Texture-Based Grading: Application to Alzheimer's Disease Detection." International Workshop on Patch-based Techniques in Medical Imaging: 82-89.

Holland, D., L. Chang, T. M. Ernst, M. Curran, S. D. Buchthal, D. Alicata, J. Skranes, H. Johansen, A. Hernandez and R. Yamakawa (2014). "Structural growth trajectories and rates of change in the first 3 months of infant brain development." JAMA neurology **71**(10): 1266-1274.

Hu, S. and D. L. Collins (2007). "Joint level-set shape modeling and appearance modeling for brain structure segmentation." NeuroImage **36**(3): 672-683.

Hu, S., P. Coupe, J. C. Pruessner and D. L. Collins (2014). "Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation." Hum Brain Mapp **35**(2): 377-395.

Huhtaniska, S., E. Jääskeläinen, T. Heikka, J. S. Moilanen, H. Lehtiniemi, J. Tohka, J. V. Manjón, P. Coupé, L. Björnholm and H. Koponen (2017). "Long-term antipsychotic and benzodiazepine use and brain volume changes in schizophrenia: The Northern Finland Birth Cohort 1966 study." Psychiatry Research: Neuroimaging.

Huttenlocher, P. R. and A. S. Dabholkar (1997). "Regional differences in synaptogenesis in human cerebral cortex." Journal of comparative Neurology **387**(2): 167-178.

Jack, C. R., Jr., R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, E. G. Tangalos and E. Kokmen (2000). "Rates of hippocampal atrophy correlate with change in clinical status in aging and AD." Neurology **55**(4): 484-489.

Jack, C. R., D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste and S. D. Weigand (2013). "Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers." The Lancet Neurology **12**(2): 207-216.

Jack, C. R., R. C. Petersen, Y. C. Xu, S. C. Waring, P. C. O'Brien, E. G. Tangalos, G. E. Smith, R. J. Ivnik and E. Kokmen (1997). "Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease." Neurology **49**(3): 786-794.

Jenkinson, M., C. F. Beckmann, T. E. Behrens, M. W. Woolrich and S. M. Smith (2012). "Fsl." Neuroimage **62**(2): 782-790.

Jung, Y., J. L. Whitwell, J. R. Duffy, E. A. Strand, M. M. Machulda, M. L. Senjem, C. R. Jack, V. J. Lowe and K. A. Josephs (2016). "Regional β-amyloid burden does not correlate with cognitive or language deficits in Alzheimer's disease presenting as aphasia." European journal of neurology **23**(2): 313-319.

Kantarci, K., C. Yang, J. A. Schneider, M. L. Senjem, D. A. Reyes, V. J. Lowe, L. L. Barnes, N. T. Aggarwal, D. A. Bennett, G. E. Smith, R. C. Petersen, C. R. Jack, Jr. and B. F. Boeve (2012). "Ante mortem amyloid imaging and beta-amyloid pathology in a case with dementia with Lewy bodies." Neurobiology of Aging **33**(5): 878-885.

Karran, E. and J. Hardy (2014). "Antiamyloid therapy for Alzheimer's disease--are we on the right road?" N Engl J Med **370**(4): 377-378.

Kerchner, G., C. Hess, K. Hammond-Rosenbluth, D. Xu, G. Rabinovici, D. Kelley, D. Vigneron, S. Nelson and B. Miller (2010). "Hippocampal CA1 apical neuropil atrophy in mild Alzheimer disease visualized with 7-T MRI." Neurology **75**(15): 1381-1387.

Kerchner, G. A., G. K. Deutsch, M. Zeineh, R. F. Dougherty, M. Saranathan and B. K. Rutt (2012). "Hippocampal CA1 apical neuropil atrophy and memory performance in Alzheimer's disease." Neuroimage **63**(1): 194-202.

Koikkalainen, J., J. Lotjonen, L. Thurfjell, D. Rueckert, G. Waldemar and H. Soininen (2011). "Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease." Neuroimage **56**(3): 1134-1144.

Koikkalainen, J., H. Rhodius-Meester, A. Tolonen, F. Barkhof, B. Tijms, A. W. Lemstra, T. Tong, R. Guerrero, A. Schuh and C. Ledig (2016). "Differential diagnosis of neurodegenerative diseases using structural MRI data." NeuroImage: Clinical **11**: 435-449.

Komlagan, M., V.-T. Ta, X. Pan, J.-P. Domenger, D. L. Collins, P. Coupé and A. s. D. N. Initiative (2014). "Anatomically constrained weak classifier fusion for early detection of Alzheimer's disease." International Workshop on Machine Learning in Medical Imaging: 141-148.

Kulaga-Yoskovitz, J., B. C. Bernhardt, S.-J. Hong, T. Mansi, K. E. Liang, A. J. Van Der Kouwe, J. Smallwood, A. Bernasconi and N. Bernasconi (2015). "Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset." Scientific data **2**.

Kumfor, F., L.-A. Sapey-Triomphe, C. E. Leyton, J. R. Burrell, J. R. Hodges and O. Piguet (2014). "Degradation of emotion processing ability in corticobasal syndrome and Alzheimer's disease." Brain **137**(11): 3061-3072.

La Joie, R., A. Perrotin, V. De La Sayette, S. Egret, L. Doeuvre, S. Belliard, F. Eustache, B. Desgranges and G. Chételat (2013). "Hippocampal subfield volumetry in mild cognitive impairment, Alzheimer's disease and semantic dementia." NeuroImage: Clinical **3**: 155-162.

Lebel, C., M. Gee, R. Camicioli, M. Wieler, W. Martin and C. Beaulieu (2012). "Diffusion tensor imaging of white matter tract evolution over the lifespan." Neuroimage **60**(1): 340-352.

Lehmann, M., P. M. Ghosh, C. Madison, R. Laforce Jr, C. Corbetta-Rastelli, M. W. Weiner, M. D. Greicius, W. W. Seeley, M. L. Gorno-Tempini and H. J. Rosen (2013). "Diverging patterns of amyloid deposition and hypometabolism in clinical variants of probable Alzheimer's disease." Brain **136**(3): 844-858.

Lenroot, R. K. and J. N. Giedd (2010). "Sex differences in the adolescent brain." Brain Cogn **72**(1): 46-55.

Lenroot, R. K., N. Gogtay, D. K. Greenstein, E. M. Wells, G. L. Wallace, L. S. Clasen, J. D. Blumenthal, J. Lerch, A. P. Zijdenbos, A. C. Evans, P. M. Thompson and J. N. Giedd (2007). "Sexual dimorphism of brain developmental trajectories during childhood and adolescence." Neuroimage **36**(4): 1065-1073.

Lerch, J. P. and A. C. Evans (2005). "Cortical thickness analysis examined through power analysis and a population simulation." Neuroimage **24**(1): 163-173.

Leung, K. K., J. Barnes, M. Modat, G. R. Ridgway, J. W. Bartlett, N. C. Fox, S. Ourselin and A. s. D. N. Initiative (2011). "Brain MAPS: an automated, accurate and robust brain extraction technique using a template library." Neuroimage **55**(3): 1091-1108.

Li, H., G. Jiang, R. Wang, J. Zhang, Z. Wang, W.-S. Zheng and B. Menze (2018). "Fully Convolutional Network Ensembles for White Matter Hyperintensities Segmentation in MR Images." arXiv preprint arXiv:1802.05203.

Liao, S., Y. Gao, J. Lian and D. Shen (2013). "Sparse patch-based label propagation for accurate prostate localization in CT images." IEEE transactions on medical imaging **32**(2): 419-434.

Lin, T.-W., Y.-F. Liu, Y.-H. Shih, S.-J. Chen, T.-Y. Huang, C.-Y. Chang, C.-H. Lien, L. Yu, S.-H. Chen and Y.-M. Kuo (2015). "Neurodegeneration in amygdala precedes hippocampus in the APPswe/PS1dE9 mouse model of Alzheimer's disease." Current Alzheimer Research **12**(10): 951-963.

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken and C. I. Sánchez (2017). "A survey on deep learning in medical image analysis." arXiv preprint arXiv:1702.05747.

Liu, M., A. Kitsch, S. Miller, V. Chau, K. Poskitt, F. Rousseau, D. Shaw and C. Studholme (2016). "Patch-based augmentation of Expectation–Maximization for brain MRI tissue segmentation at arbitrary age after premature birth." NeuroImage **127**: 387-408.

Liu, M., D. Zhang, D. Shen and A. s. D. N. Initiative (2012). "Ensemble sparse classification of Alzheimer's disease." NeuroImage **60**(2): 1106-1116.

Liu, Y., G. Spulber, K. K. Lehtimäki, M. Könönen, I. Hallikainen, H. Gröhn, M. Kivipelto, M. Hallikainen, R. Vanninen and H. Soininen (2011). "Diffusion tensor imaging and tract-based spatial statistics in Alzheimer's disease and mild cognitive impairment." Neurobiology of aging **32**(9): 1558-1571.

Lorente de Nó, R. (1934). "Studies on the structure of the cerebral cortex. II. Continuation of the study of the ammonic system." Journal für Psychologie und Neurologie.

Lotjonen, J., R. Wolz, J. Koikkalainen, V. Julkunen, L. Thurfjell, R. Lundqvist, G. Waldemar, H. Soininen and D. Rueckert (2011). "Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease." Neuroimage **56**(1): 185-196.

Lotjonen, J. M., R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen and D. Rueckert (2010). "Fast and robust multi-atlas segmentation of brain magnetic resonance images." NeuroImage **49**(3): 2352-2365.

Makropoulos, A., P. Aljabar, R. Wright, B. Huning, N. Merchant, T. Arichi, N. Tusor, J. V. Hajnal, A. D. Edwards, S. J. Counsell and D. Rueckert (2016). "Regional growth and atlasing of the developing human brain." Neuroimage **125**: 456-478.

Mangin, J.-F., D. Riviere, A. Cachia, E. Duchesnay, Y. Cointepas, D. Papadopoulos-Orfanos, P. Scifo, T. Ochiai, F. Brunelle and J. Regis (2004). "A framework to study the cortical folding patterns." Neuroimage **23**: S129-S138.

Mangin, J. F., M. Chupin, M. O. Habert, H. Benali, B. Batrancourt, O. Colliot and S. Lehericy (2014). "CATI: harmonisation de l'imagerie dans les maladies neurodégénératives." Revue Neurologique **170**: A222.

Manjon, J. V. and P. Coupe (2016). "volBrain: An Online MRI Brain Volumetry System." Front Neuroinform **10**: 30.

Manjon, J. V., P. Coupe and A. Buades (2015). "MRI noise estimation and denoising using non-local PCA." Med Image Anal **22**(1): 35-47.

Manjon, J. V., P. Coupe, A. Buades, D. L. Collins and M. Robles (2010a). "MRI superresolution using self-similarity and image priors." Int J Biomed Imaging **2010**: 425891.

Manjon, J. V., P. Coupe, A. Buades, V. Fonov, D. Louis Collins and M. Robles (2010b). "Non-local MRI upsampling." Medical image analysis **14**(6): 784-792.

Manjon, J. V., P. Coupe, A. Buades, D. Louis Collins and M. Robles (2012). "New methods for MRI denoising based on sparseness and self-similarity." Med Image Anal **16**(1): 18-27.

Manjon, J. V., P. Coupe, L. Concha, A. Buades, D. L. Collins and M. Robles (2013). "Diffusion weighted image denoising using overcomplete local PCA." PLoS One **8**(9): e73021.

Manjon, J. V., P. Coupe, L. Marti-Bonmati, D. L. Collins and M. Robles (2010c). "Adaptive non-local means denoising of MR images with spatially varying noise levels." J Magn Reson Imaging **31**(1): 192-203.

Manjón, J. V., P. Coupé, P. Raniga, Y. Xia, J. Fripp and O. Salvado (2016). "HIST: HyperIntensity Segmentation Tool." International Workshop on Patch-based Techniques in Medical Imaging: 92-99.

Manjon, J. V., S. F. Eskildsen, P. Coupe, J. E. Romero, D. L. Collins and M. Robles (2014). "Nonlocal intracranial cavity extraction." Int J Biomed Imaging **2014**: 820205.

Manjón, J. V., J. Tohka and M. Robles (2010). "Improved estimates of partial volume coefficients from noisy brain MRI using spatial context." Neuroimage **53**(2): 480-490.

Manjunath, B. S. and W.-Y. Ma (1996). "Texture features for browsing and retrieval of image data." IEEE Transactions on pattern analysis and machine intelligence **18**(8): 837-842.

Mazziotta, J. C., A. W. Toga, A. Evans, P. Fox and J. Lancaster (1995). "A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM)." NeuroImage **2**(2): 89-101.

McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price and E. M. Stadlan (1984). "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology **34**(7): 939-944.

Medland, S. E., N. Jahanshad, B. M. Neale and P. M. Thompson (2014). "Whole-genome analyses of whole-brain data: working within an expanded search space." Nature neuroscience **17**(6): 791-800.

Miller, M. I., L. Younes, J. T. Ratnanather, T. Brown, H. Trinh, E. Postell, D. S. Lee, M.-C. Wang, S. Mori and R. O'Brien (2013). "The diffeomorphometry of temporal lobe structures in preclinical Alzheimer's disease." NeuroImage: Clinical **3**: 352-360.

Mills, K. L., A.-L. Goddings, M. M. Herting, R. Meuwese, S.-J. Blakemore, E. A. Crone, R. E. Dahl, B. Güroğlu, A. Raznahan and E. R. Sowell (2016). "Structural brain development between childhood and adulthood: Convergence across four longitudinal samples." NeuroImage **141**: 273-281.

Morey, R. A., E. S. Selgrade, H. R. Wagner, S. A. Huettel, L. Wang and G. McCarthy (2010). "Scan–rescan reliability of subcortical brain volumes derived from automated segmentation." Human brain mapping **31**(11): 1751-1762.

Moroso, A., A. Ruet, D. Lamargue-Hamel, F. Munsch, M. Deloire, P. Coupé, J. Charré-Morin, A. Saubusse, J.-C. Ouallet and V. Planche (2017). "Microstructural analyses of the posterior cerebellar lobules in relapsing-onset multiple sclerosis and their implication in cognitive impairment." PloS one **12**(8): e0182479.

Morra, J., Z. Tu, A. Toga and P. Thompson (2008). "Automatic segmentation of MS lesions using a contextual model for the MICCAI grand challenge." Multiple Sclerosis Lesion Segmentation Challenge: 1-7.

Mueller, S. G., M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga and L. Beckett (2005). "The Alzheimer's disease neuroimaging initiative." Neuroimaging Clinics of North America **15**(4): 869-877.

Müller, M. J., D. Greverus, P. R. Dellani, C. Weibrich, P. R. Wille, A. Scheurich, P. Stoeter and A. Fellgiebel (2005). "Functional implications of hippocampal volume and diffusivity in mild cognitive impairment." Neuroimage **28**(4): 1033-1042.

Næss-Schmidt, E., A. Tietze, J. U. Blicher, M. Petersen, I. K. Mikkelsen, P. Coupé, J. V. Manjón and S. F. Eskildsen (2016). "Automatic thalamus and hippocampus segmentation from MP2RAGE: comparison of publicly available methods and implications for DTI quantification." International journal of computer assisted radiology and surgery **11**(11): 1979-1991.

Nestor, S. M., R. Rupsingh, M. Borrie, M. Smith, V. Accomazzi, J. L. Wells, J. Fogarty, R. Bartha and A. s. D. N. Initiative (2008). "Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database." Brain **131**(9): 2443-2454.

Nir, T. M., N. Jahanshad, J. E. Villalon-Reina, A. W. Toga, C. R. Jack, M. W. Weiner, P. M. Thompson and A. s. D. N. Initiative (2013). "Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging." NeuroImage: clinical **3**: 180-195.

O'Dwyer, L., F. Lamberton, A. L. Bokde, M. Ewers, Y. O. Faluyi, C. Tanner, B. Mazoyer, D. O'Neill, M. Bartley and D. R. Collins (2012). "Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment." PloS one **7**(2): e32441.

Ohm, T. and H. Braak (1987). "Olfactory bulb changes in Alzheimer's disease." Acta neuropathologica **73**(4): 365-369.

Ostby, Y., C. K. Tamnes, A. M. Fjell, L. T. Westlye, P. Due-Tonnessen and K. B. Walhovd (2009). "Heterogeneity in subcortical brain development: A structural magnetic resonance imaging study of brain maturation from 8 to 30 years." J Neurosci **29**(38): 11772-11782.

Park, M. T. M., J. Pipitone, L. H. Baer, J. L. Winterburn, Y. Shah, S. Chavez, M. M. Schira, N. J. Lobaugh, J. P. Lerch and A. N. Voineskos (2014). "Derivation of high-resolution MRI atlases of the human cerebellum at 3T and segmentation using multiple automatically generated templates." Neuroimage **95**: 217-231.

Patenaude, B., S. M. Smith, D. N. Kennedy and M. Jenkinson (2011). "A Bayesian model of shape and appearance for subcortical brain segmentation." Neuroimage **56**(3): 907-922.

Paxinos, G. and J. K. Mai (2004). The human nervous system, Academic Press.

Pfefferbaum, A., T. Rohlfing, M. J. Rosenbloom, W. Chu, I. M. Colrain and E. V. Sullivan (2013). "Variation in longitudinal trajectories of regional brain volumes of healthy men and women (ages 10 to 85 years) measured with atlas-based parcellation of MRI." Neuroimage **65**: 176-193.

Phelps, E. A. (2004). "Human emotion and memory: interactions of the amygdala and hippocampal complex." Current opinion in neurobiology **14**(2): 198-202.

Phelps, E. A. and J. E. LeDoux (2005). "Contributions of the amygdala to emotion processing: from animal models to human behavior." Neuron **48**(2): 175-187.

Pipitone, J., M. T. M. Park, J. Winterburn, T. A. Lett, J. P. Lerch, J. C. Pruessner, M. Lepage, A. N. Voineskos, M. M. Chakravarty and A. s. D. N. Initiative (2014). "Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates." Neuroimage **101**: 494-512.

Planche, V., A. Ruet, P. Coupe, D. Lamargue-Hamel, M. Deloire, B. Pereira, J. V. Manjon, F. Munsch, N. Moscufo, D. S. Meier, C. R. Guttmann, V. Dousset, B. Brochet and T. Tourdias (2017). "Hippocampal microstructural damage correlates with memory impairment in clinically isolated syndrome suggestive of multiple sclerosis." Multiple Sclerosis Journal: 1352458516675750.

Poldrack, R. A. and K. J. Gorgolewski (2014). "Making big data open: data sharing in neuroimaging." Nat Neurosci **17**(11): 1510-1517.

Potvin, O., A. Mouiha, L. Dieumegarde, S. Duchesne and I. Alzheimer's Disease Neuroimaging (2016). "Normative data for subcortical regional volumes over the lifetime of the adult human brain." Neuroimage **137**: 9-20.

Poulin, S. P., R. Dautoff, J. C. Morris, L. F. Barrett, B. C. Dickerson and A. s. D. N. Initiative (2011). "Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity." Psychiatry Research: Neuroimaging **194**(1): 7-13.

Prasad, G., S. H. Joshi, T. M. Nir, A. W. Toga, P. M. Thompson and A. s. D. N. Initiative (2015). "Brain connectivity and novel network measures for Alzheimer's disease classification." Neurobiology of aging **36**: S121-S131.

Prima, S., S. Ourselin and N. Ayache (2002). "Computation of the mid-sagittal plane in 3-D brain images." IEEE Transactions on Medical Imaging **21**(2): 122-138.

Prince, M., R. Bryce and C. Ferri (2011). World Alzheimer Report 2011: The benefits of early diagnosis and intervention, Alzheimer's Disease International.

Protter, M., M. Elad, H. Takeda and P. Milanfar (2009). "Generalizing the nonlocal-means to super-resolution reconstruction." IEEE Trans Image Process **18**(1): 36-51.

Pruessner, J. C., L. M. Li, W. Serles, M. Pruessner, D. L. Collins, N. Kabani, S. Lupien and A. C. Evans (2000). "Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories." Cereb Cortex **10**(4): 433-442.

Qiu, A., C. Fennema-Notestine, A. M. Dale, M. I. Miller and I. Alzheimer's Disease Neuroimaging (2009). "Regional shape abnormalities in mild cognitive impairment and Alzheimer's disease." Neuroimage **45**(3): 656-661.

Raznahan, A., P. Shaw, F. Lalonde, M. Stockman, G. L. Wallace, D. Greenstein, L. Clasen, N. Gogtay and J. N. Giedd (2011). "How does your cortex grow?" Journal of Neuroscience **31**(19): 7174-7177.

Rohlfing, T., R. Brandt, R. Menzel and C. R. Maurer, Jr. (2004). "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains." NeuroImage **21**(4): 1428-1442.

Romero, J. E., P. Coupe, R. Giraud, V. T. Ta, V. Fonov, M. T. Park, M. M. Chakravarty, A. N. Voineskos and J. V. Manjon (2017a). "CERES: A new cerebellum lobule segmentation method." Neuroimage **147**: 916-924.

Romero, J. E., P. Coupe and J. V. Manjón (2016). "High Resolution Hippocampus Subfield Segmentation Using Multispectral Multiatlas Patch-Based Label Fusion." International Workshop on Patch-based Techniques in Medical Imaging: 117-124.

Romero, J. E., P. Coupé and J. V. Manjón (2017b). "HIPS: A new hippocampus subfield segmentation method." NeuroImage **163**: 286-295.

Romero, J. E., J. V. Manjon, J. Tohka, P. Coupe and M. Robles (2015). "NABS: non-local automatic brain hemisphere segmentation." Magn Reson Imaging **33**(4): 474-484.

Ronneberger, O., P. Fischer and T. Brox (2015). "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention: 234-241.

Rose, S. E., L. Andrew and J. B. Chalk (2008). "Gray and white matter changes in Alzheimer's disease: a diffusion tensor imaging study." Journal of Magnetic Resonance Imaging **27**(1): 20-26.

Sanroma, G., V. Andrea, O. M. Benkarim, J. V. Manjón, P. Coupé, O. Camara, G. Piella and M. A. González Ballester (2017). "Early Prediction of Alzheimer's Disease with Non-local Patch-Based Longitudinal Descriptors." Patch-Based Techniques in Medical Imaging:: 74-81.

Scherrer, B., F. Forbes, C. Garbay and M. Dojat (2009). "Distributed local MRF models for tissue and structure brain segmentation." IEEE Trans Med Imaging **28**(8): 1278-1295.

Shan, L., C. Zach, C. Charles and M. Niethammer (2014). "Automatic atlas-based three-label cartilage segmentation from MR knee images." Medical image analysis **18**(7): 1233-1246.

Shaw, P., N. J. Kabani, J. P. Lerch, K. Eckstrand, R. Lenroot, N. Gogtay, D. Greenstein, L. Clasen, A. Evans and J. L. Rapoport (2008). "Neurodevelopmental trajectories of the human cerebral cortex." Journal of Neuroscience **28**(14): 3586-3594.

Shen, D., S. Moffat, S. M. Resnick and C. Davatzikos (2002). "Measuring size and shape of the hippocampus in MR images using a deformable shape model." NeuroImage **15**(2): 422-434.

Shen, D., G. Wu and H.-I. Suk (2017). "Deep learning in medical image analysis." Annual Review of Biomedical Engineering(0).

Sherif, T., P. Rioux, M.-E. Rousseau, N. Kassis, N. Beck, R. Adalat, S. Das, T. Glatard and A. C. Evans (2015). "CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research." Recent Advances and the Future Generation of Neuroinformatics Infrastructure: 102.

Shiee, N., P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi and D. L. Pham (2010). "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions." NeuroImage **49**(2): 1524-1535.

Siadat, M. R., H. Soltanian-Zadeh and K. V. Elisevich (2007). "Knowledge-based localization of hippocampus in human brain MRI." Comput Biol Med **37**(9): 1342-1360.

Smith, S. M. (2002). "Fast robust automated brain extraction." Human brain mapping **17**(3): 143-155.

Sørensen, L., C. Igel, A. Pai, I. Balas, C. Anker, M. Lillholm, M. Nielsen and A. s. D. N. Initiative (2017). "Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry." NeuroImage: Clinical **13**: 470-482.

Souplet, J.-C., C. Lebrun, N. Ayache and G. Malandain (2008). An automatic segmentation of T2-FLAIR multiple sclerosis lesions. The MIDAS Journal-MS Lesion Segmentation (MICCAI 2008 Workshop).

Sowell, E. R., B. S. Peterson, P. M. Thompson, S. E. Welcome, A. L. Henkenius and A. W. Toga (2003). "Mapping cortical change across the human life span." Nat Neurosci **6**(3): 309-315.

Spalding, K. L., O. Bergmann, K. Alkass, S. Bernard, M. Salehpour, H. B. Huttner, E. Bostrom, I. Westerlund, C. Vial, B. A. Buchholz, G. Possnert, D. C. Mash, H. Druid and J. Frisen (2013). "Dynamics of hippocampal neurogenesis in adult humans." Cell **153**(6): 1219-1227.

St-Jean, S., P. Coupe and M. Descoteaux (2016). "Non Local Spatial and Angular Matching: Enabling higher spatial resolution diffusion MRI datasets through adaptive denoising." Med Image Anal **32**: 115-130.

Stiles, J. and T. L. Jernigan (2010). "The basics of brain development." Neuropsychology review **20**(4): 327-348.

Stonnington, C. M., C. Chu, S. Klöppel, C. R. Jack, J. Ashburner, R. S. J. Frackowiak and I. Alzheimer Disease Neuroimaging (2010). "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease." Neuroimage **51**(4): 1405-1413.

Styner, M., J. Lee, B. Chin, M. Chin, O. Commowick, H. Tran, S. Markovic-Plese, V. Jewells and S. Warfield (2008). "3D segmentation in the clinic: A grand challenge II: MS lesion segmentation." Midas Journal **2008**: 1-6.

Suk, H.-I., S.-W. Lee, D. Shen and A. s. D. N. Initiative (2017). "Deep ensemble learning of sparse regression models for brain disease diagnosis." Medical image analysis **37**: 101-113.

Sutour, C., C.-A. Deledalle and J.-F. Aujol (2014). "Adaptive regularization of the NL-means: Application to image and video denoising." IEEE Transactions on image processing **23**(8): 3506-3521.

Ta, V. T., R. Giraud, D. L. Collins and P. Coupe (2014). "Optimized patchMatch for near real time and accurate label fusion." Med Image Comput Comput Assist Interv **17**(Pt 3): 105-112.

Tabert, M. H., X. Liu, R. L. Doty, M. Serby, D. Zamora, G. H. Pelton, K. Marder, M. W. Albers, Y. Stern and D. Devanand (2005). "A 10-item smell identification scale related to risk for Alzheimer's disease." Annals of neurology **58**(1): 155-160.

Tanskanen, P., J. M. Veijola, U. K. Piippo, M. Haapea, J. A. Miettunen, J. Pyhtinen, E. T. Bullmore, P. B. Jones and M. K. Isohanni (2005). "Hippocampus and amygdala volumes in schizophrenia and other psychoses in the Northern Finland 1966 birth cohort." Schizophr Res **75**(2-3): 283-294.

Tellouck, L., M. Durieux, P. Coupe, A. Cougnard-Gregoire, J. Tellouck, T. Tourdias, F. Munsch, A. Garrigues, C. Helmer, F. Malet, J. F. Dartigues, V. Dousset, C. Delcourt and C. Schweitzer (2016). "Optic Radiations Microstructural Changes in Glaucoma and Association With Severity: A Study Using 3Tesla-Magnetic Resonance Diffusion Tensor Imaging." Investigative ophthalmology & visual science **57**(15): 6539-6547.

Tomas-Fernandez, X. and S. K. Warfield (2011). "A new classifier feature space for an improved multiple sclerosis lesion segmentation." Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on: 1492-1495.

Tong, T., Q. Gao, R. Guerrero, C. Ledig, L. Chen, D. Rueckert and A. s. D. N. Initiative (2017a). "A Novel Grading Biomarker for the Prediction of Conversion From Mild Cognitive Impairment to Alzheimer's Disease." IEEE Transactions on Biomedical Engineering **64**(1): 155-165.

Tong, T., C. Ledig, R. Guerrero, A. Schuh, J. Koikkalainen, A. Tolonen, H. Rhodius, F. Barkhof, B. Tijms and A. W. Lemstra (2017b). "Five-class Differential Diagnostics of Neurodegenerative Diseases using Random Undersampling Boosting." NeuroImage: Clinical **15**: 613-624.

Tong, T., R. Wolz, P. Coupe, J. V. Hajnal, D. Rueckert and I. Alzheimer's Disease Neuroimaging (2013). "Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling." Neuroimage **76**: 11-23.

Tong, T., R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert and A. s. D. N. Initiative (2014). "Multiple instance learning for classification of dementia in brain MRI." Medical image analysis **18**(5): 808-818.

Tong, T., R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J. V. Hajnal and D. Rueckert (2015). "Discriminative dictionary learning for abdominal multi-organ segmentation." Medical image analysis **23**(1): 92-104.

Tottenham, N. and M. A. Sheridan (2009). "A review of adversity, the amygdala and the hippocampus: a consideration of developmental timing." Front Hum Neurosci **3**: 68.

Trujillo-Estrada, L., J. C. Dávila, E. Sánchez-Mejias, R. Sánchez-Varo, A. Gomez-Arboledas, M. Vizuete, J. Vitorica and A. Gutiérrez (2014). "Early neuronal loss and axonal/presynaptic damage is associated with accelerated amyloid-β accumulation in AβPP/PS1 Alzheimer's disease mice subiculum." Journal of Alzheimer's Disease **42**(2): 521-541.

Tustison, N. J., B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich and J. C. Gee (2010). "N4ITK: improved N3 bias correction." IEEE Trans Med Imaging **29**(6): 1310-1320.

Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer and M. Joliot (2002). "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain." Neuroimage **15**(1): 273-289.

Valverde, S., M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver and X. Lladó (2017). "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach." NeuroImage **155**: 159-168.

Van Horn, J. D. and A. W. Toga (2014). "Human neuroimaging as a “Big Data” science." Brain imaging and behavior **8**(2): 323-331.

van Praag, H., A. F. Schinder, B. R. Christie, N. Toni, T. D. Palmer and F. H. Gage (2002). "Functional neurogenesis in the adult hippocampus." Nature **415**(6875): 1030-1034.

Vijayakumar, N., N. B. Allen, G. Youssef, M. Dennison, M. Yucel, J. G. Simmons and S. Whittle (2016). "Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume." Hum Brain Mapp **37**(6): 2027-2038.

Villemagne, V. L., K. E. Pike, G. Chetelat, K. A. Ellis, R. S. Mulligan, P. Bourgeat, U. Ackermann, G. Jones, C. Szoeke, O. Salvado, R. Martins, G. O'Keefe, C. A. Mathis, W. E. Klunk, D. Ames, C. L. Masters and C. C. Rowe (2011). "Longitudinal assessment of Abeta and cognition in aging and Alzheimer disease." Annals of Neurology **69**(1): 181-192.

Wachinger, C., M. Brennan, G. C. Sharp and P. Golland (2017a). "Efficient Descriptor-Based Segmentation of Parotid Glands With Nonlocal Means." IEEE Transactions on Biomedical Engineering **64**(7): 1492-1502.

Wachinger, C., M. Reuter and T. Klein (2017b). "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy." Neuroimage.

Walhovd, K. B., A. M. Fjell, J. Giedd, A. M. Dale and T. T. Brown (2016). "Through Thick and Thin: a Need to Reconcile Contradictory Results on Trajectories in Human Cortical Development." Cerebral Cortex: bhv301.

Walhovd, K. B., L. T. Westlye, I. Amlien, T. Espeseth, I. Reinvang, N. Raz, I. Agartz, D. H. Salat, D. N. Greve, B. Fischl, A. M. Dale and A. M. Fjell (2011). "Consistent neuroanatomical age-related volume differences across multiple samples." Neurobiol Aging **32**(5): 916-932.

Wang, H., B. Avants and P. Yushkevich (2012). A combined joint label fusion and corrective learning approach. MICCAI Workshop on Multi-Atlas Labeling.

Wang, H., S. R. Das, J. W. Suh, M. Altinay, J. Pluta, C. Craige, B. Avants, P. A. Yushkevich and I. Alzheimer's Disease Neuroimaging (2011). "A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation." Neuroimage **55**(3): 968-985.

Wang, H., J. W. Suh, S. R. Das, J. B. Pluta, C. Craige and P. A. Yushkevich (2013a). "Multi-atlas segmentation with joint label fusion." IEEE transactions on pattern analysis and machine intelligence **35**(3): 611-623.

Wang, H. and P. A. Yushkevich (2013). "Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation." Frontiers in neuroinformatics **7**.

Wang, L., K. C. Chen, Y. Gao, F. Shi, S. Liao, G. Li, S. G. Shen, J. Yan, P. K. Lee and B. Chow (2014a). "Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization." Medical physics **41**(4).

Wang, L., F. Shi, G. Li, Y. Gao, W. Lin, J. H. Gilmore and D. Shen (2014b). "Segmentation of neonatal brain MR images using patch-driven level sets." NeuroImage **84**: 141-158.

Wang, Z., A. C. Bovik, H. R. Sheikh and E. P. Simoncelli (2004). "Image quality assessment: From error visibility to structural similarity." IEEE Transactions on Image Processing **13**(4): 600-612.

Wang, Z., C. Donoghue and D. Rueckert (2013b). "Patch-based segmentation without registration: application to knee MRI." International Workshop on Machine Learning in Medical Imaging: 98-105.

Wang, Z., J. Wang, H. Zhang, R. Mchugh, X. Sun, K. Li and Q. X. Yang (2015). "Interhemispheric functional and structural disconnection in Alzheimer's disease: a combined resting-state fMRI and DTI study." PLoS One **10**(5): e0126310.

Wee, C.-Y., P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang and D. Shen (2012). "Identification of MCI individuals using structural and functional connectivity networks." Neuroimage **59**(3): 2045-2056.

Weier, K., V. Fonov, K. Lavoie, J. Doyon and D. L. Collins (2014). "Rapid automatic segmentation of the human cerebellum and its lobules (RASCAL)—Implementation and application of the patch-based label-fusion technique with a template library to segment the human cerebellum." Human brain mapping **35**(10): 5026-5039.

Wells, W. M., W. E. L. Grimson, R. Kikinis and F. A. Jolesz (1996). "Adaptive segmentation of MRI data." IEEE transactions on medical imaging **15**(4): 429-442.

Wiest-Daessle, N., S. Prima, P. Coupe, S. P. Morrissey and C. Barillot (2007). "Non-local means variants for denoising of diffusion-weighted and diffusion tensor MRI." Med Image Comput Comput Assist Interv **10**(Pt 2): 344-351.

Wiest-Daessle, N., S. Prima, P. Coupe, S. P. Morrissey and C. Barillot (2008). "Rician noise removal by non-Local Means filtering for low signal-to-noise ratio MRI: applications to DT-MRI." Med Image Comput Comput Assist Interv **11**(Pt 2): 171-179.

Winterburn, J. L., J. C. Pruessner, S. Chavez, M. M. Schira, N. J. Lobaugh, A. N. Voineskos and M. M. Chakravarty (2013). "A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging." Neuroimage **74**: 254-265.

Wolz, R., P. Aljabar, J. V. Hajnal, A. Hammers, D. Rueckert and A. s. D. N. Initiative (2010). "LEAP: learning embeddings for atlas propagation." NeuroImage **49**(2): 1316-1325.

Wolz, R., P. Aljabar, J. V. Hajnal, J. Lotjonen and D. Rueckert (2011a). "Manifold learning combining imaging with non-imaging information." Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on: 1637-1640.

Wolz, R., C. Chu, K. Misawa, M. Fujiwara, K. Mori and D. Rueckert (2013). "Automated abdominal multi-organ segmentation with subject-specific atlas generation." IEEE transactions on medical imaging **32**(9): 1723-1730.

Wolz, R., V. Julkunen, J. Koikkalainen, E. Niskanen, D. P. Zhang, D. Rueckert, H. Soininen and J. Lotjonen (2011b). "Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease." PLoS One **6**(10): e25446.

Wu, D., C. Ceritoglu, M. I. Miller and S. Mori (2016a). "Direct estimation of patient attributes from anatomical MRI based on multi-atlas voting." NeuroImage: Clinical **12**: 570-581.

Wu, G., P. Coupé, Y. Zhan, B. Munsell and D. Rueckert (2015a). Patch-Based Techniques in Medical Imaging: First International Workshop, Patch-MI 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 9, 2015, Revised Selected Papers, Springer.

Wu, G., P. Coupé, Y. Zhan, B. C. Munsell and D. Rueckert (2016b). Patch-Based Techniques in Medical Imaging: Second International Workshop, Patch-MI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings, Springer.

Wu, G., M. Kim, G. Sanroma, Q. Wang, B. C. Munsell, D. Shen and A. s. D. N. Initiative (2015b). "Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition." NeuroImage **106**: 34-46.

Wu, G., B. C. Munsell, Y. Zhan, W. Bai, G. Sanroma and P. Coupé (2017). Patch-Based Techniques in Medical Imaging: Third International Workshop, Patch-MI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Proceedings, Springer.

Wu, G., Q. Wang, D. Zhang, F. Nie, H. Huang and D. Shen (2014). "A generative probability model of joint label fusion for multi-atlas based brain segmentation." Medical image analysis **18**(6): 881-890.

Xu, Y., T. Géraud, É. Puybareau, I. Bloch and J. Chazalon (2017). "White Matter Hyperintensities Segmentation in a Few Seconds Using Fully Convolutional Network and Transfer Learning." 501-514.

Yang, X., P. Rossi, A. Jani, T. Ogunleye, W. Curran and T. Liu (2015). "3D Prostate Segmentation in Ultrasound Images Using Patch-Based Anatomical Feature." Medical physics **42**(6): 3685-3685.

Yushkevich, P. A., R. S. Amaral, J. C. Augustinack, A. R. Bender, J. D. Bernstein, M. Boccardi, M. Bocchetta, A. C. Burggren, V. A. Carr and M. M. Chakravarty (2015a). "Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: towards a harmonized segmentation protocol." Neuroimage **111**: 526-541.

Yushkevich, P. A., J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee and G. Gerig (2006). "User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability." Neuroimage **31**(3): 1116-1128.

Yushkevich, P. A., J. B. Pluta, H. Wang, L. Xie, S. L. Ding, E. C. Gertje, L. Mancuso, D. Kliot, S. R. Das and D. A. Wolk (2015b). "Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment." Human brain mapping **36**(1): 258-287.

Zandifar, A., V. Fonov, P. Coupe, J. Pruessner, D. L. Collins and I. Alzheimer's Disease Neuroimaging (2017). "A comparison of accurate automatic hippocampal segmentation methods." Neuroimage **155**: 383-393.

Zhang, D., D. Shen and I. Alzheimer's Disease Neuroimaging (2012). "Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers." PloS one **7**(3): e33182.

Zhao, L., U. Ruotsalainen, J. Hirvonen, J. Hietala and J. Tohka (2010). "Automatic cerebral and cerebellar hemisphere segmentation in 3D MRI: adaptive disconnection algorithm." Medical image analysis **14**(3): 360-372.

Zhou, J. and J. C. Rajapakse (2005). "Segmentation of subcortical brain structures using fuzzy templates." NeuroImage **28**(4): 915-924.

Zhuang, X. and J. Shen (2016). "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI." Medical image analysis **31**: 77-87.

Ziegler, G., R. Dahnke, L. Jancke, R. A. Yotter, A. May and C. Gaser (2012). "Brain structural trajectories over the adult lifespan." Hum Brain Mapp **33**(10): 2377-2389.

Zijdenbos, A. P., B. M. Dawant, R. A. Margolin and A. C. Palmer (1994). "Morphometric analysis of white matter lesions in MR images: method and validation." IEEE Trans Med Imaging **13**(4): 716-724.