E-infrastructure, cortical mesh segmentation, quality control environment : a red thread for neuroscientists

Alberto Redolfi

université PARIS-SACLAY

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

NNT : 2017SACLS191

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
L'UNIVERSITE PARIS-SUD

ECOLE DOCTORALE N°575

EOBE | Electrical, optical, bio-physics and engineering

Spécialité de doctorat : Imagerie et Physique Médicale

Par

M. Alberto Redolfi

E-infrastructure, segmentation du cortex, environnement de contrôle qualité : un fil rouge pour les neuroscientifiques

Thèse présentée et soutenue à Gif-sur-Yvette, le 12 Juillet 2017 :

Composition du Jury :

M. Fabrice Crivello	Directeur de recherche	CNRS-CEA-Université de Bordeaux	Présidente du Jury
M. François Rousseau	Professeur	Institut Mines-Telecom, Telecom Bretagne, INSERM, Université de Strasbourg-CNRS	Rapporteur
M. Olivier Coulon	Directeur de recherche	Institut de Neurosciences de la Timone, Aix Marseille Université, CNRS	Rapporteur
M. Giovanni B Frisoni	Professeur	Université de Genève, Hôpitaux universitaires de Genève, IRCCS Centro San Giovanni di Dio Fatebenefratelli, Brescia	Examinateur
M. Édouard Duchesnay	Chargé de recherche	CATI, Neurospin, CEA, Paris-Saclay University	Examinateur
M. Jean-François Mangin	Directeur de recherche	CATI, Neurospin, CEA, Paris-Saclay University	Directeur de thèse

Contents

Chapter 1

Introduction

GENERAL INTRODUCTION

Evolution of brain imaging in neurodegenerative diseases

Brain imaging was regarded as an elective examination in patients with cognitive decline 25 years ago [1].

The practice parameters for diagnosis and evaluation of dementia defined by the American Academy of Neurology considered magnetic resonance (MR) as an 'optional' assessment [2,3].

Over time, imaging in dementia has moved from a negative, exclusionary role, to one able to add positive diagnostic and prognostic information.

In the late 1990s, the traditional exclusionary approach was abandoned in favour of an inclusive one [4,5]. Rapid advances in neuroimaging technologies such as PET, single photon emission CT, MR spectroscopy, diffusion tensor imaging (DTI), and functional MRI have offered new vision into the pathophysiology of brain diseases and, in particular, of the Alzheimer's disease (AD) [6]. Consequently, new powerful data-analysis methods have been developed [7].

Since the beginning of the 21st century, the development of innovative techniques for cortical thickness measurement, region-of-interest (ROI) based volumetry, automated voxel-based morphometry, and multivariate statistics have emerged [7–9].

With the development of novel image processing techniques, the complexity of neuroimaging analysis has started to increase significantly: images with higher spatial resolution and acquired with longer time scans yielded to a greater amount of voxels to be processed.

At the same time, the more computationally intensive algorithms needed required more computational resources.

Due to their high costs of purchase and maintenance, several research and medical imaging facilities have not yet been able to afford the necessary equipment and tools to satisfy the computational demand of the most advanced neuroimaging analysis.

Under these circumstances, during the last decade the neuroimaging community has started to develop distributed e-infrastructures, which combine high-performance computing and innovative customizable algorithms together with large databases and image datasets, collected worldwide on a day-to-day basis [10–13] (Figure 1).

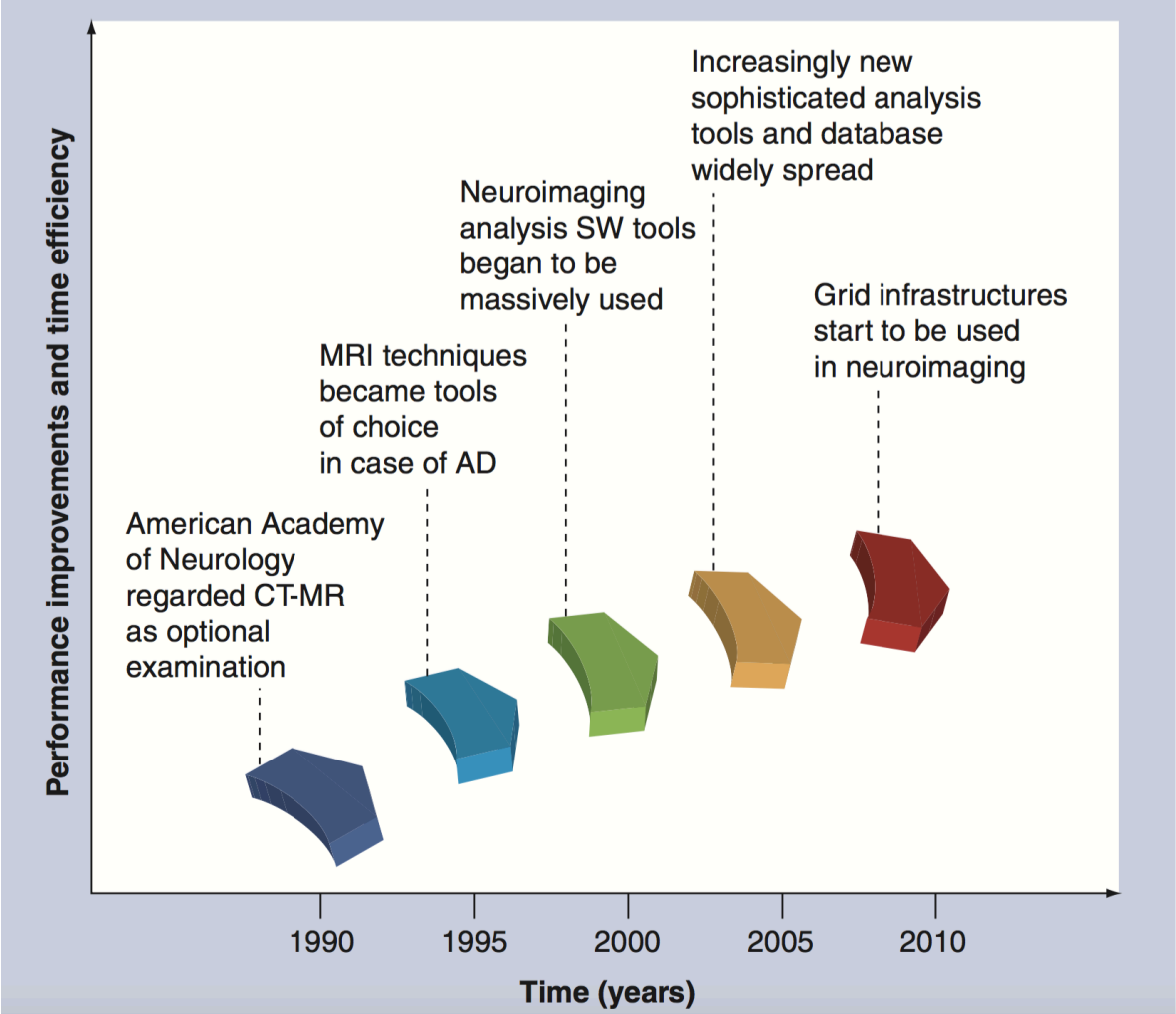


Figure 1 - Evolutionary steps of neuroimaging analysis tools from early 1990s to the actual exploitation of e-infrastructures. AD: Alzheimer's disease; CT-MR: Computed tomography-magnetic resonance; SW: Software.

Neuroimaging in the era of big-data

The maturation of *in vivo* neuroimaging has led to a deluge of digital information about the human brain [14]. Brain imaging research studies are performed to explore the brain in action or at rest, to study how it is built and wired, as well as what happens when things degenerate. As MRI technology have improved, extended, and become faster, so too have the amount of brain data generated. Once proven robust, researchers have started to adopt these enhanced methodologies – doubling or tripling the volume of data per subject. Today, neuroscientists routinely collect a larger amount of study data in a single day than was put together in over a year just a decade ago.

To get the glimpse of this trend, confirming an evidence already published by John Darrell Van

Horn et al. in 2014, a simple examination of fMRI articles from the NeuroImage journal indicates that since 1995 the amount of data collected has doubled roughly every 30 months (Figure 2).

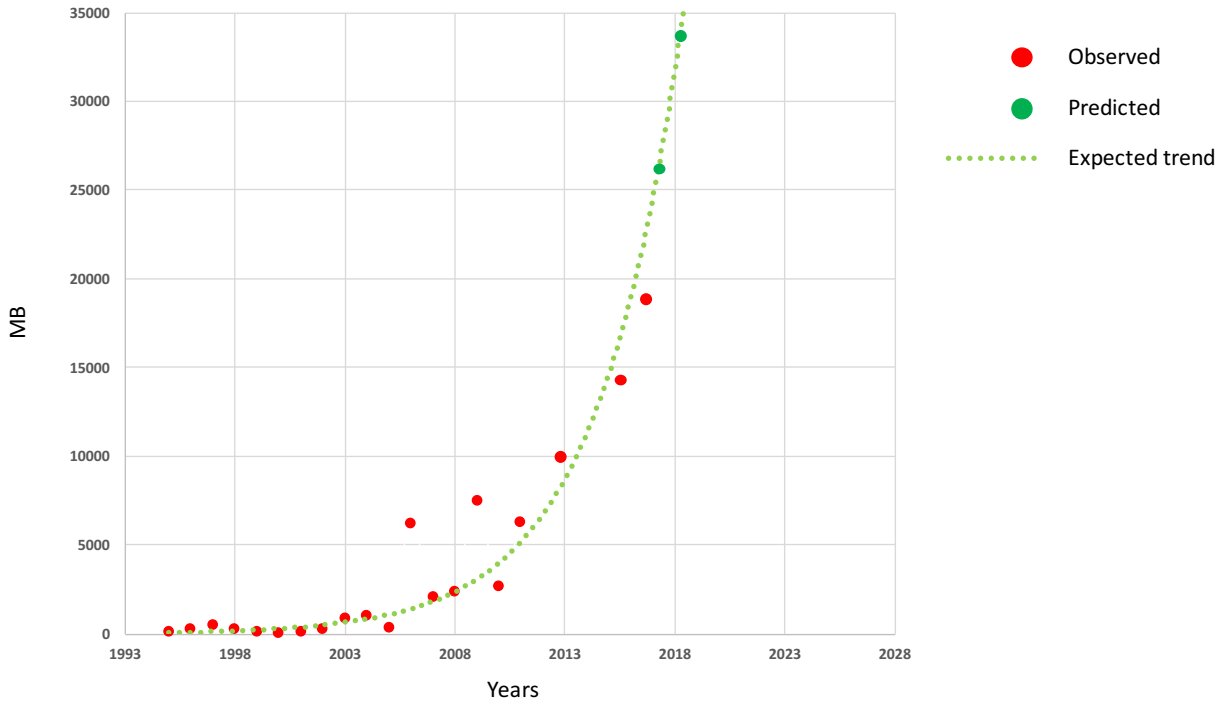


Figure 2 - The amount of acquired MRI data reported from published articles in the NeuroImage journal has doubled every 30 months showing an exponential trend in the upcoming years.

At this rate, by 2018 the amount of acquired neuroimaging data alone, discounting header information and before the additional files generated during data processing and statistical analysis, may exceed an average of 30 GB per published research study. This is an under-estimation for raw dataset sizes since, as noted above, advances in MRI physics are accelerating the pace at which brain data can be acquired per unit time. Therefore, it is safe to say that human neuroimaging is now officially a "big data" science.

The neuroimaging e-infrastructures and services

A deep change in the research paradigm has started to be experienced (Figure 3).

Figure 3 - Paradigm shift from "*ex-vivo*" to "*in-vivo*" and "*in-silico*".

Individual desktop computers are now no longer suitable to analyze terabyte and potentially petabytes worth of brain. The growth, availability and accessibility of imaging has therefore led to the development of computational e-infrastructures, which offer neuroscientists access to large and well-curated image databases, services such as sophisticated image analysis algorithm pipelines and powerful computational resources, as well as three-dimensional visualization and statistical tools. At present, only a few imaging laboratories have the technical expertise and computational resources required to merge multiple large data sets and explore scientific questions relating to larger populations.

In Europe and North America, e-infrastructures are being developed to fill the gap between data acquisition and information extraction. The initiatives described in the following bullet-points and summarized in Table 1 share the common vision of offering a full range of imaging services to neuroscientists. Access to such novel platforms can be provided through web browsers, science gateway portals, or via Linux command line interfaces. The range of databases and algorithms is markedly variable, and computational resources are based on either a central server or cluster or a distributed grid infrastructure.

Presently, we are in the very early days of public services for computational neuroscience, and the current infrastructures might undergo substantial reshaping in the near future. However, it is relevant to note what is available today, as these infrastructures can be to neuroscientists

what the Large Hadron Collider is to physicists – the framework where the most 'muscular' experiments can be run and audacious hypotheses can be tested.

	VIP	CATI	EMIF-AD	GAAIN	HBP-MIP	CBRAIN	LONI	neuGRID
HW features								
Infrastructure typology	Grid	HPC	Grid	HPC & distributed local resources	Grid, HPC, & distributed local resources	Grid & HPC	HPC	Grid/Cloud
Storage capacity	5 PB	0.5 PB	5 TB	5 TB	200 TB + HPC distributed storage	80 TB + 1 PB HPC resources	12 PB	25 TB
Core resources	15000 CPU cores	500 CPU cores	500 CPU cores	200 CPU cores	1000 CPU cores + HPC PRACE cores	80000 CPU cores + Judge/Juropa/Kisti HPC cores	10000 CPU cores	500 CPU cores + EGI distributed cores
SW features								
Platform Link	https://www.creatis.insa-lyon.fr/vip/	http://cati-neuroimaging.com/	https://emif.custodix.com/transmart	http://www.gaaindata.org	https://mip.humanbrainproject.eu	https://portal.cbrain.mcgill.ca/login	http://pipeline.loni.usc.edu	https://neugrid4you.eu
Image processing algorithms	For structural and functional analysis	For structural, functional ,and diffusion analysis	For structural analysis	For structural, functional and diffusion analysis	For structural analysis	For structural, functional and diffusion analysis	For structural, functional and diffusion analysis	For structural, functional and diffusion analysis
Pre-processing tools	Freesurfer and FSL data preparation functions	Soma-workflow, AIMS, Python libraries	Freesurfer and FSL data preparation functions	Loni-Pipeline pre-processing modules via Docker containers	Morphobox, SPM	MINC libraries, conversion tools	AFNI, AIR, ITK, MINC, MIND, Utilities, Shape tools	MINC libraries, mricron, DCMTK, ITK, VTK
Post-processing tools	Freesurfer and FSL core functions, MRtrix3, France-Life imaging tools	BrainVisa suite, SPM, Freesurfer	Freesurfer and FSL core functions, LEAP, SPM	Loni-Pipeline post-processing modules via Docker containers	Machine Learning tools for clusterization and classification via Docker containers	CLASP-CIVET; FSL; Freesurfer	SPM, FSL, Freesurfer, MIND, LONI DTI suite, GAMMA, SVPASEG, WAIR; Bioinformatics suite	AFNI, Freesurfer, FSL, BrainVisa, SPM, UMCP, FCP-CPAC, Adaboost, MRTRIX, LEAP, GDI segmenter
QC	//	Manual quality assurance (i.e.: SnapBase)	Manual quality assurance (i.e.: Enigma QC tool)	//	//	Manual quality assurance (i.e.: Civet_QC_tool)	//	//
Statistical tools	//	Python libraries	R plugins	SOCR, R libraries	R libraries	R, RMINC, SurfStat packages	SOCR, SVT, MAST libraries	R, Octave, and Python libraries
Image datasets provided								
Community	//	AD, PD, LBD, FTD	AD	AD	AD	//	AD, PD, PSY, PSTD	AD, LBD, WMD, PSY
Modality	//	T13D, R-fMRI, DTI, PET, ASL	T13D	T13D, R-fMRI, DTI, PET, ASL	T13D	//	T13D, R-fMRI, DTI, PET, ASL	T13D, R-fMRI, DTI, PET
Dataset	//	Memento; Memento-Vascod; Memento-Amyging	1000-AD-cohort	NACC; WLS; WRAP; ADNI; I-ADNI; E-ADNI; LMRR; DIAN; AIBL; OASIS; ARWIBO; LAADC; DART; NIAGADS; DESCRIPA; CAMD-CODR; BIOCARD; ACE; BHR; FNAD; HRS; HELIAD; EDSD; CLSA; FHS; INDD	ADNI1	//	ABIDE, ADNI1-GO-2-3; AIBL, PAD/CRYO, PPMI, DOD	OASIS; MIRIAD; ADNI1-GO-2; I-ADNI; H2H; ADHD-200; 1000 FCP; E-ADNI; ARWIBO; EDSD; MAGNIMS; fBIRN phase I & II; NUSDAST; ABIDE; INDI; COBRE
Mechanism of provision	//	Data-Sharing	Data-Federation	Data-Federation	Data-Federation	//	Data-Sharing	Data Sharing
Accessibility	//	Restricted-Access	Restricted-Access	Facilitated-Access	Open-Access	//	Open-Access	Open-Access; Facilitated-Access; Restricted-Access

Table 1 – Largest e-infrastructures developed worldwide.

HW: hardware; SW: software; PB: petabyte; TB: terabyte; HPC: high performance computing; CPU: central processing unit; QC: Quality Control; AD: Alzheimer's disease; WMD: white matter disease; PD: Parkinson's disease; FTD: frontotemporal dementia: LBD: Lewy body dementia; PSY: psychiatric disease; T13D: volumetric sequence weighted in T1; PET: positron emission tomography; R-fMRI: resting functional MRI; DTI: diffusion tensor imaging; ASL: arterial spin labeling.

ADNI: the Alzheimer's Disease Neuroimaging Initiative comprises different phases: ADNI1, ADNIGO, ADNI2, and ADNI3; AIBL: The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; NACC: National Alzheimer's Coordinating Centre; ARWIBO: Alzheimer's Disease Repository Without Borders; HELIAD: Hellenic Longitudinal Investigation on Aging and Diet; DART: Dementia and Aging Research of Taiwan;

CAMD-CODR: Coalition Against Major Diseases; E-ADNI: European ADNI (it is also known as PharmaCOG); CLSA: Canadian Longitudinal Study on Aging; DIAN: Dementia Inherited Alzheimer Network; WLS: Wisconsin Longitudinal Study; WRAP: Wisconsin Registry for Alzheimer's Prevention; HRS: Health and Retirement Study; FNAD: French National Alzheimer Database; I-ADNI: Italian ADNI; NIAGDS: NIA Genetics of Alzheimer's Disease Data Storage Site; Brain Health Registry; ACE: Foundacio' ACE; BIOCARD: Predictors of Cognitive Decline Among Normal Individuals; LMRR: Laboratory of Magnetic Resonance Research; DESCRIPA: Development of screening guidelines and criteria for predementia Alzheimer's disease; FHS: Framingham Heart Study; LAADC: Layton Aging & Alzheimer's Disease Center; INDD: Integrated Neurodegenerative Disease Database. PPMI: Parkinson's Progression Markers Initiative; DOD: Department of Defense, a study of brain raging in Vietnam war veterans; MIRIAD: Minimal Interval Resonance Imaging in Alzheimer's Disease; H2H comparison study: Head to head comparison study between Civet and Freesufer; ADHD-200: Attention Deficit Hyperactivity Disorder; MAGNIMS: Magnetic Resonance Imaging in MS; ABIDE: Autism Brain Imaging Data Exchange; PAD/CRYO: Public Anonymized Dataset/Cryosection; OASIS: Open Access Series of Imaging Studies; 1000-FCP: 1000 functional connectomes project; INDI: International Neuroimaging Data Sharing Initiative project; EDSD: European diffusion tensor imaging study in dementia; FBIRN: The Functional Bioinformatics Research Network, it is composed by many phases (phase-I and II have been archived in neuGRID); NUSDAST: Northwestern University Schizophrenia Data and Software Tool; COBRE: dataset of the Center for Biomedical Research Excellence.

Open-Access: means downloading and analyzing data free and without restrictions or formalities; Facilitated-Access: the analysis is possible only after obtaining an account from data-owners and permission for the intended use. The Support Centre of the platform (i.e.: GAAIN and neuGRID) facilitates data exploitation by connecting the neuroscientist requesting data access with the PI of the research dataset; Restricted-Access: access is restricted exclusively to co-workers or collaborators of the initiative. Authorization from other external users to access these data must be obtained from the owner by submitting an application and generally signing an *ad hoc* scientific agreement.

- VIP (Virtual Imaging Platform):

  VIP is a web-portal for medical simulation and image data analysis. VIP can be used by neuroscientists worldwide without specific knowledge of programming or other IT skills beyond the use of a web browser. VIP relies on European Grid Infrastructure's (EGI) High-Throughput Compute and Online Storage resources made available through the Biomed Virtual Organization. As of January 2017, VIP has 990 registered users, with 115 users active during the last quarter of 2016. On average, researchers using VIP for their work consume about 30 years of CPU time per month [16].

- CATI (Centre pour l'Acquisition et le Traitement de l'Image):

  CATI is a service platform to provide assistance for acquiring, analyzing, organizing, and sharing neuroimaging data among scientific and medical communities, supporting

numerous studies in the fields of neurodegenerative diseases, psychiatry, and therapeutic trials (about 40 different studies are supported). CATI includes more than 10'000 subjects often acquired longitudinally with MRI and PET. The CATI initiative offers a complete portfolio of image processing tools, including international standards like voxel-based and tract-based morphometry, as well as distributed database services. Via these services, the CATI initiative mutualizes the resources and offer valid support through experts at the scale of big data. The tools and services of CATI have been developed in line with the ADNI standards [17].

- EMIF-AD (European Medical Information Framework for Alzheimer's Disease):
  EMIF is an initiative that aims to re-use existing electronic health records (EHR). The EMIF platform allows researchers to browse and process harmonized patients-level data to support research. A data repository known as TranSMART provides a common data schema for a variety of data type (e.g., clinical, socio-demographical, neuropsychological, biological, imaging) coming from different European datasets, such as: Antwerp, E-ADNI, Lausanne, EDAR-VUMC, DESCRIPA, CITA, IDIBAPS, Leuven, Gothenburg. This is called 1000-AD-Cohort study and it is composed by 195 AD, 515 MCI, 383 healthy elderly control. Volumetric T13D scans can be processed to determine which imaging biomarker (or which combination) offers the best diagnostic or prognostic outcomes [18].

- GAAIN (Global Alzheimer's Association Interactive Network):
  At the base of the GAAIN initiative is the idea that research efforts could be vastly expanded in scope and capabilities if data were linked to a global infrastructure that would enable scientists to access and use interlinked repositories of data on thousands of participants at risk of Alzheimer's disease. GAAIN created and maintains a network of interested parties worldwide (the GAAIN's Data Partners) to develop, launch, and sustain a distributed database infrastructure. GAAIN platform federates 452'000 subjects (mainly AD, MCI, CTR). GAAIN comprises a GAAIN central server, located at the University of South California (USC), and a Data Partner Client (DPC) that runs at each Data Partner's location. The GAAIN central server manages communication between the DPCs and the central server, requesting data from DPC's as needed to fulfill queries launched from the GAAIN server's GUI interface (the so called

"Interrogator"). The DPC contains a light-weight web server, a database, and a "transformer" that automatically maps the Partner's data into the GAAIN common data model schema. The DPC controls communication between the Partner site and the GAAIN server. In this way, the Interrogator of the GAAIN platform provides a highly visible and interactive platform for investigating data aggregated from Partner's that meet the users search criteria on the biggest AD ever [19-21].

- HBP - MIP (Human Brain Project – Medical Informatics Platform):
  The HBP IT architecture is unique, utilizing cloud-based collaboration and development platforms with databases, workflow systems, petabyte storage, and super-computers (i.e.: Barcelona Supercomputing Centre - BSC -, the Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Universita` e Ricerca - CINECA -, the Centro Svizzero di Calcolo Scientifico - CSCS -, and the Julich Supercomputing Centre - JSC -). Specifically, the Medical Informatics Platform (MIP) of the HBP includes federation nodes in different European hospitals for *in situ* querying of anonymized clinical data and data integration. Indeed, hospitals and other medical databases contain vast amounts of data about health and disease patients that represent an enormous asset to researchers. MIP supports neuroscientists to generate new knowledge by providing a way of organizing and using big data without the need to physically transfer data. The MIP integrates heterogeneous data and federates them into harmonized databases with *ad hoc* interfaces for navigation, data mining, and machine learning (ML) investigations. The MIP final goal is to develop effective pipelines for extracting biological signatures of diseases from multi-level data [22].

- CBRAIN (Canadian Brain Imaging Research Platform):
  CBRAIN is a network of Canada's nine leading brain imaging research centers linked within a platform for distributed processing and data sharing. The CBRAIN platform addresses issues of advanced networking, transparent access to remote computer resources, integration of heterogeneous environments, tool usability, and web-based three-dimensional visualization, by providing users with a comprehensive collaborative web portal enabling them to manage, transfer, share, analyze and visualize their imaging data. Because of its distributed nature and ease of use, the CBRAIN platform connects Canadian brain imaging research centers not only to two High Performance

Computing (HPC) centers (i.e., Judge/Jurope in Germany; KISTI in Korea) but also to multiple collaborating sites around the world. Researchers can launch their jobs through an easy-to-use web interface, and allow the platform to handle data transfers, job scheduling on Grid or HPC, and results. CBRAIN is based on a set of leased high-speed wide area network links, i.e.: CAnet [23].

- LONI (Laboratory of Neuro Imaging):
  LONI focuses on the development of image analysis methods and their application in health as well as in neurological and psychiatric disorders. LONI hosts, among others, the ADNI-1 dataset, which comprises clinical and genetic information as well as scans from 872 older people with mild cognitive impairment (MCI), 342 people with AD, and 417 healthy elders (CTR). All of them have been followed with high-resolution structural MRI. In addition, in the context of the ADNI-GO and ADNI-2 initiatives, where 850 new subjects (150 CTR, 500 MCI, 200 AD) were recruited, fMRI, DTI, and ASL data were acquired, as well as biannual 18F-fluorodeoxyglucose PET (FDG-PET) and amyloid PET. Within the forthcoming ADNI-3 initiative, patients will also be studied with 18F-AV-1451 to understand the mechanism of Tau protein deposition for a time frame of 5 years. Algorithms for data analysis are accessible both independently and through the graphical LONI Pipeline, a user-friendly workflow management system. The LONI Pipeline enables automated measurement of functional and morphometric analyses, dynamic assessment of volume, shape (e.g., curvature) and form (e.g., thickness) features, as well as the extraction and association between cognitive, genetic, clinical, and behavioral biomarkers. For external investigators, LONI provides access to a large HPC infrastructure, physically located at USC, for computationally intensive image analyses [24].

- NeuGRID (Grid-based e-infrastructure for data archiving/communication and computationally intensive applications in the medical sciences):
  The neuGRID platform makes use of grid and computing services across four main nodes, i.e., GNUBILA (Argonay, France), NeuroSpin (Paris, France), HUG (Geneva, Switzerland) and IRCCS Fatebenefratelli (Brescia, Italy). NeuGRID has been developed with the final aim of overcoming the hurdles that the average scientist meets when trying to set up advanced experiments in computational neuroimaging, thereby

empowering a larger base of neuroscientists. The final production version was completed in March 2015. Although originally built for neuroscientists working in the field of AD, the infrastructure has been expanded to other medical fields, such as white matter diseases and psychiatric disorders. NeuGRID currently hosts 17 datasets comprising 1'122 AD, 1'779 MCI, 2'896 CTR, 158 Multiple Sclerosis patients, and 2'478 psychiatric disorders (e.g., autism, schizophrenia, ADHD) overall. NeuGRID is compliant with international standards for data collection, data management, and grid abstraction [25]. As of April 2017, neuGRID has 275 registered users, of which 31 active during the first quarter of 2017.

Research on imaging biomarkers: the case of cortical thickness

Thanks to its safety and accessibility, magnetic resonance imaging (MRI) is extensively used in research field, largely contributing to our understanding of the pathophysiology of neurodegenerative disorders of the brain.

Early diagnosis and development of effective disease-modifying drugs in AD would be facilitated by the availability of an accurate (sensitive and specific cross-sectionally and prospectively) disease marker.

Brain atrophy and cortical thinning represent the most obvious disease marker for AD. Atrophy is the most evident change, detectable in the brain either at autopsy or with brain imaging techniques.

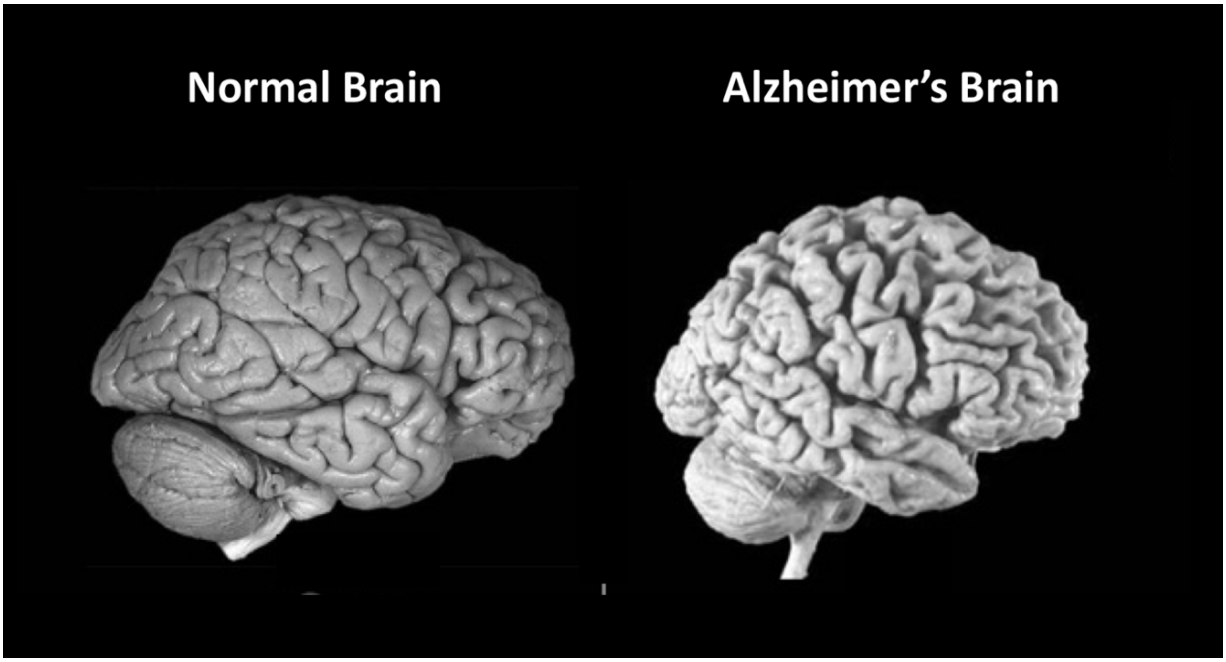


Figure 4 - Brain atrophy is a disease marker of Alzheimer's disease and is well suited to be used to assess the efficacy of drugs aimed at slowing or halting neurodegeneration

Figure 4 shows the preparations of a normal brain and a pathological brain with Alzheimer's, where the latter shows gross shrinkage of the superficial layer of the brain consisting of grey matter.

MRI has the sufficient spatial accuracy and resolution to detect the subtle atrophic changes that take place in the early Alzheimer's brain. However, the digital images resulting from acquisition must be processed with sophisticated algorithms that treat the intensity of the biological signal in each voxel (the basic constituting image unit, a virtual cube of less than 1 mm^3) taking into account the intensity of surrounding voxels with high demanding and computationally intensive iterative algorithms.

In the last twenty years, many pipelines have started to appear to assess the cortical thickness in a repeatable and quantifiable manner. So far there is not a clear picture of which performs the best. Therefore, although this would be an ideal disease marker for early diagnosis and drug development in AD, cortical thickness needs much larger validation that it presently has [26].

Last but not least, it needs to be stressed that the available algorithms can take as long as 35 hours per brain to run on a state-of-the-art workstation. In this scenario, the exploitation of e-infrastructures is very much appropriate.

The neuroscientists are hard to please

Despite the consideration on the numbers of e-infrastructures and services listed above, an educated eye can realize a point of possible further development, that is, an automated quality control (QC) specifically drawn for high-throughput image analysis.

The QC procedure aims to check that the image-processing algorithm produced the expected result. Should something go wrong, the algorithm needs to be re-run on those specific images after having optimized certain parameters.

Unfortunately, no tool able judge the goodness of terabyte of imaging output results is yet available on the market, ready to be incorporated in these e-infrastructures. Today, an experts need to visually inspect the generated outputs, carrying out the QC on an individual basis.

Some instances of QC procedures for high throughput image analyses are presently starting to appear in the biomedical analysis community, but these are still at a very early stage of development and, most importantly, still require a direct control of experts.

Among other delivered products, this thesis represents a leap forward in the QC of big-data

image analysis by providing an automated, customizable, and scalable environment (Table 2).

Automatic	Extraction and comparison of learned patterns and features allowing intelligent decision-making through a supervised learning approach
Pipeline specific	Suitable for the most used cortical thickness algorithms
Disease invariant	Flexible and robust to accommodate different stages of disease (e.g.: AD, MCI, CTR) during the evaluation process
Customizable	Set QCE out according to different cortical thickness algorithm requirements (file formats, conventions, etc.)
Scalable	Capable to handle growing amounts of outputs

Table 2 – Neuroscientists wishful features for an innovative QC environment.

AIMS AND OUTLINE

This thesis is focused on the current and future potentialities of the brain e-infrastructures, on the new challenges related to the availability and accessibility of big-data, the variability of imaging biomarkers estimated by similar pipelines, the need for a fully automated quality control environment (QCE), and the effort to design an innovative machine learning tool helping neuroscientists in their daily and data-intensive research activity.

The research described in the following chapters is based on specific research questions:

- *Which kind of analyses can be triggered in the big brain science era?*
  Aim of Chapter 2 was to investigate the multimodal and multiscale pattern of the data archived in the e-infrastructures and outline their usage for advanced brain (hyper)models development. We provided an overview of the available e-infrastructures and considered how computational neuroscience in neurodegenerative disease might evolve further on. Such experimental environments will be instrumental to the success of ambitious scientific initiatives with high societal impact, such as the prevention of Alzheimer disease.

- *What is a real life use-case today in such e-infrastructures?*
  Chapter 3 focuses on the head-to-head comparison of two popular pipelines (Civet and Freesurfer) for the estimation of the cortical thickness in CTR, sMCI, pMCI and AD subjects. Cortical thinning is a recognized marker of neurodegeneration, a putative marker of disease progression, and a reasonable surrogate outcome in clinical trials. That represents a typical scenario of interest for a neuroscientist willing to assess and investigate the sensitivity associated to structural changes of the cortical mantle of the two well-known algorithms and to investigate the amount of variance between them.

- *When and how a neuroscientist will significantly benefit of an automated Quality Control Environment?*
  Indeed, the works presented in Chapter 2 and 3 led to a bottleneck: the unresolved issue of manual quality control of pipelines' outputs. Aim of Chapter 4 was to design an automated and effective QCE tailored to the needs of neuroscientists. A supervised Machine Learning (ML) was implemented to discriminate "good" versus "bad" 3D

cortical meshes of both Civet and Freesurfer pipelines. Furthermore, a multilabel classification approach was developed to help neuroscientists locate the cortical segmentation artefacts.

REFERENCES

1. van Straaten ECW, Scheltens P, Barkhof F: MRI and CT in the diagnosis of vascular dementia. J. Neurol. Sci. 226, 9–12 (2004).
2. Practice parameter for diagnosis and evaluation of dementia. Report of the quality standards subcommittee of the American Academy of Neurology. Neurology 44, 2203–2206 (1994).
3. Frisoni GB, Scheltens P, Galuzzi S et al.: Neuroimaging tools to rate regional atrophy, subcortical cerebrovascular, and regional cerebral blood flow and metabolism: consensus paper of the EADC. J. Neurol. Neurosurg. Psychiatr. 74, 1371–1381 (2003).
4. Blow N: Neuroscience tools: brain insights. Nat. Methods 5(11), 981–987 (2008).
5. Scheltens P, Fox N, Barkhof F et al.: Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. Lancet Neurol. 1(1), 13–21 (2002).
6. Knopman DS, DeKosky ST, Cummings JL et al.: Practice parameter: Diagnosis of dementia (an evidence-based review): report of the quality standards subcommittee of the American Academy of Neurology. Neurology 56, 1143–1155 (2001).
7. DeCarli C: The role of neuroimaging in dementia. Clin. Geriatr. Med. 17(2), 255–279 (2001).
8. Teipel SJ, Meindl T, Grinberg L et al.: Novel MRI techniques in the assessment of dementia. Eur. J. Nuc. Med. Mol. Imaging 35, S58–S69 (2008).
9. Ashburner J, Csernansky JG, Davatzikos C et al.: Computer-assisted imaging to assess brain structure in healthy and diseased brains. Lancet Neurol. 2, 79–88 (2003).
10. Glatard T, Lingrand D, Montagnat J et al.: Impact of the execution context on Grid job performances. Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid, IEEE Computer Society, 713–718 (2007).
11. Germain C, Breton V, Clarysse P et al.: Grid-enabling medical image analysis. J. Clin. Monit. Comput. 19(4–5), 339–349 (2005).
12. Foster I: The Grid: computing without bounds. Sci. Am. 288(4), 78–85 (2003).
13. Redolfi A, McClatchey R, Anjum A, Zijdenbos A, Manset D, Barkhof F, Spenger C, Legré Y, Wahlund LO, Barattieri di San Pietro C, Frisoni GB. Grid infrastructures for computational neuroscience: the neuGRID example. Future Neurology. 2009, Vol. 4, No. 6, Pages 703-722, doi:10.2217/fnl.09.53.
14. Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. Brain Imaging Behav. 2014 Jun;8(2):323-31. doi: 10.1007/s11682-013-9255-y.
15. Frisoni GB, Redolfi A, Manset D, Rousseau MÉ, Toga A, Evans AC. Virtual imaging laboratories for marker discovery in neurodegenerative diseases. Nat Rev Neurol. 2011 Jul 5;7(8):429-38. doi: 10.1038/nrneurol.2011.99.
16. Glatard T, Lartizien C, Gibaud B, da Silva RF, Forestier G, Cervenansky F, Alessandrini M, Benoit-Cattin H, Bernard O, Camarasu-Pop S, Cerezo N, Clarysse P, Gaignard A, Hugonnard P, Liebgott H, Marache S, Marion A, Montagnat J, Tabary J, Friboulet D. A virtual imaging platform for multi-modality medical image simulation. IEEE Trans Med Imaging. 2013 Jan;32(1):110-8.
17. Operto G, Chupin M, Batrancourt B, Habert MO, Colliot O, Benali H, Poupon C, Champseix C, Delmaire C, Marie S, Rivière D, Pélégrini-Issac M, Perlbarg V, Trebossen R, Bottlaender M, Frouin V, Grigis A, Orfanos DP, Dary H, Fillon L, Azouani C, Bouyahia A, Fischer C, Edward L, Bouin M, Thoprakarn U,

Li J, Makkaoui L, Poret S, Dufouil C, Bouteloup V, Chételat G, Dubois B, Lehéricy S, Mangin JF, Cointepas Y; CATI Consortium. CATI: A Large Distributed Infrastructure for the Neuroimaging of Cohorts. Neuroinformatics. 2016 Jul;14(3):253-64. doi: 10.1007/s12021-016-9295-8.

18. Visser, Pieter Jelle et al. A EUROPEAN MEDICAL INFORMATION FRAMEWORK FOR ALZHEIMER'S DISEASE (EMIF-AD). Alzheimer's & Dementia: The Journal of the Alzheimer's Association , Volume 10 , Issue 4 , P799.
19. Ashish N, Bhatt P, Toga AW. Global Data Sharing in Alzheimer Disease Research. Alzheimer Dis Assoc Disord. 2016 Apr-Jun;30(2):160-8. doi: 10.1097/WAD.0000000000000121.
20. Toga AW, Neu SC, Bhatt P, Crawford KL, Ashish N. The Global Alzheimer's Association Interactive Network. Alzheimers Dement. 2016 Jan;12(1):49-54. doi: 10.1016/j.jalz.2015.06.1896. Epub 2015 Aug 28.
21. Neu SC, Crawford KL, Toga AW. Sharing data in the global alzheimer's association interactive network. Neuroimage. 2016 Jan 1;124(Pt B):1168-74. doi: 10.1016/j.neuroimage.2015.05.082. Epub 2015 Jun 4.
22. Amunts K, Ebell C, Muller J, Telefont M, Knoll A, Lippert T. The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain. Neuron. 2016 Nov 2;92(3):574-581. doi: 10.1016/j.neuron.2016.10.046.
23. Sherif T, Rioux P, Rousseau ME, Kassis N, Beck N, Adalat R, Das S, Glatard T, Evans AC. CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. Front Neuroinform. 2014 May 21;8:54. doi: 10.3389/fninf.2014.00054. eCollection 2014.
24. Dinov ID, Van Horn JD, Lozev KM, Magsipoc R, Petrosyan P, Liu Z, Mackenzie-Graham A, Eggert P, Parker DS, Toga AW. Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline. Front Neuroinform. 2009 Jul 20;3:22. doi: 10.3389/neuro.11.022.2009.
25. Redolfi A, Bosco P, Manset D, Frisoni GB; neuGRID consortium. Brain investigation and brain conceptualization. Funct Neurol. 2013 Jul-Sep;28(3):175-90. doi: 10.11138/FNeur/2013.28.3.175.
26. Frisoni GB, Fox NC, Jack CR Jr, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol. 2010 Feb;6(2):67-77. doi: 10.1038/nrneurol.2009.215.

Chapter 2

Brain investigation and brain conceptualization

Alberto Redolfi
Paolo Bosco
David Manset
Giovanni B Frisoni
The neuGRID consortium

Functional Neurology
2013 Jul-Sep;28(3):175-90.

ABSTRACT

The brain of a patient with Alzheimer's disease (AD) undergoes changes starting many years before the development of the first clinical symptoms. The recent availability of large prospective datasets makes it possible to create sophisticated brain models of healthy subjects and patients with AD, showing pathophysiological changes occurring over time. However, these models are still inadequate; representations are mainly single-scale and they do not account for the complexity and interdependence of brain changes. Brain changes in Alzheimer's patients occur at different levels and for different reasons: at the molecular level, changes are due to amyloid deposition; at cellular level, to loss of neuron synapses, and at tissue level, to connectivity disruption. All cause extensive atrophy of the whole brain organ. Initiatives aiming to model the whole human brain have been launched in Europe and the US with the goal of reducing the burden of brain diseases. In this work, we describe a new approach to earlier diagnosis based on a multimodal and multiscale brain concept, built upon existing and well-characterized single modalities.

INTRODUCTION

The diagnosis of neurodegenerative diseases such as Alzheimer's disease (AD) is changing dramatically. For the first time in 27 years, experts have proposed a major change in the criteria, making it possible to diagnose and treat AD earlier. The new guidelines [49] state that new instrumental technologies can be used to detect the disease even before evident memory problems or other symptoms arise.

For the first time, diagnosis aims at identifying the disease as it is developing, using results from multiple biomarker tests, such as fluorodeoxyglucose (FDG) brain scans, magnetic resonance imaging (MRI) scans and spinal taps, able to reveal indicative signs of brain degeneration. These biomarkers have been developed and tested only recently [66, 38, 54], which explains why none had been previously approved for AD diagnosis. One of the newest, the positron emission tomography (PET) scan, shows the brain plaques peculiar to the pathology of AD [30] . Others, such as cerebrospinal fluid (CSF) or MRI analyses [7, 36], provide strong indications of AD even when patients do not show any sign of dementia or memory loss.

The dynamic changes in AD biomarkers are known to occur non-linearly. Dynamic models of various neuroimaging biomarkers over time (as the disease progresses) have recently been well characterized [19], whereas genetics in combination with imaging biomarkers will soon provide even more diagnostic and prognostic information [58]. Nevertheless, there is still no multimodal and multiscale approach integrating all the information captured by each single methodology.

The new proposed criteria for AD have already started advocating the multimodal use of brain imaging techniques to examine the inner structure and function of the brain using one biological [e.g., beta-amyloid (Aβ42) or tau protein in CSF] and three imaging markers (e.g., PET amyloid imaging, ^{18}FDG PET and MRI). A number of validation studies have already been conducted, showing that these new criteria have excellent sensitivity, specificity and accuracy [33]. Although the AD scientific community welcomed the new criteria, they have still not been fully adopted in daily practice [21].

In this article, we discuss the strengths and weaknesses of the single-model approach so far used to diagnose (and prognosticate) AD, and we try to define the key issues involved in designing a brain hypermodel, leveraging on the information available at atomic, molecular, cellular and tissue level, with the aim of launching a new multimodal

approach to the study of the brain.

This article begins by focusing on the biological marker measurements used to monitor the evolution of AD, before going on to define the problems and limitations of a single-scale, single-modality approach, and finally leading the reader towards a more detailed definition of an initial brain hypermodel based on Bayesian inference and the e-infrastructures needed to implement it.

BIOLOGICAL MARKERS FOR BRAIN INVESTIGATION, DIAGNOSIS, AND MONITORING

Although it is now clear that AD involves gradual neuron failure, why this happens is still not clear. Experts believe that AD, like other chronic conditions, is the intricate result of multiple factors, rather than of a dominant cause [4]. Both age and genetics have been identified as the most common risk factors [39]. Evidence from cases where the disease runs in families with an autosomal dominant (FAD) mode of transmission indicates that the affected genes – mainly presenilin-1 (*PS1)*, presenilin-2 (*PS2)*, and amyloid precursor protein (*APP)* – are involved in the metabolism of beta-amyloid (Aβ) [60], a small protein of 40 to 42 amino acids (Fig. 1).

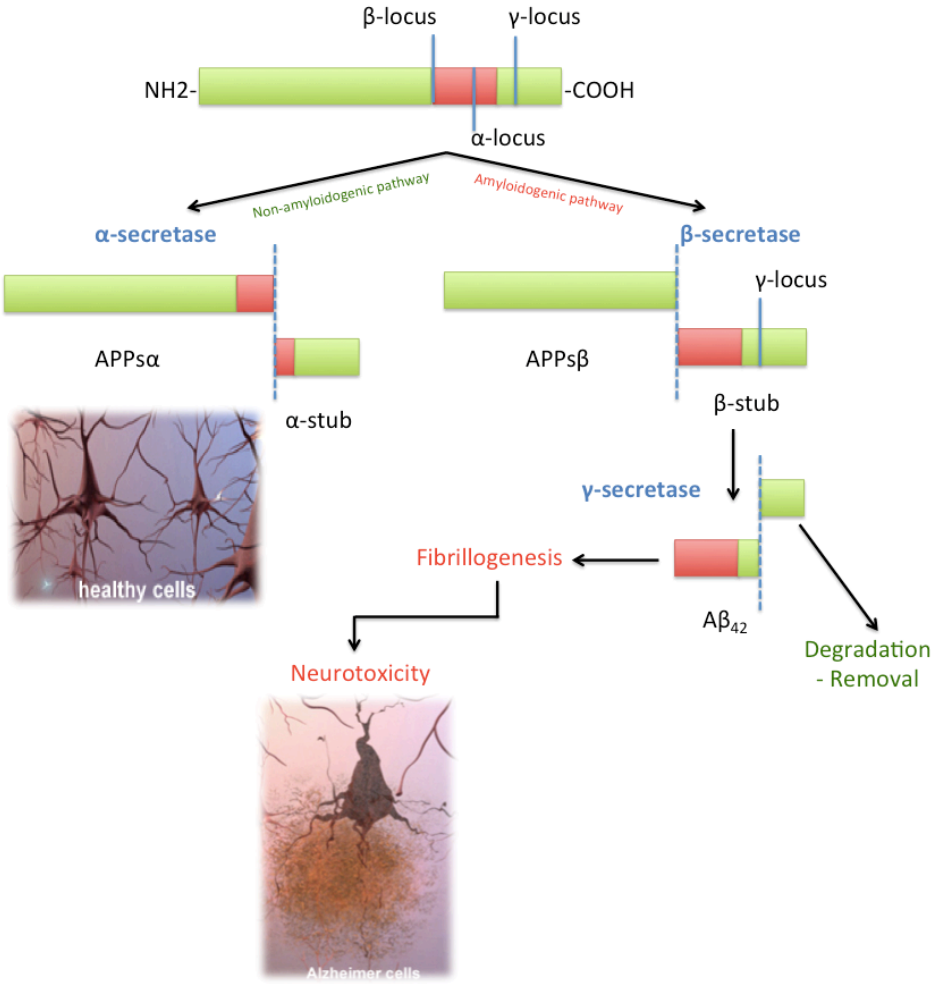


Figure 1 – The amyloid cascade. Evidence indicates that the amyloid precursor protein (APP) is involved in the metabolism of beta-amyloid (Aβ). The proteolytic processing of APP unfolds through two alternative pathways. In the non-amyloidogenic pathway, APP is processed by an α-secretase. In the amyloidogenic pathway, APP is first cleaved at a β-secretase site and subsequently cleaved by a γ-secretase complex to release the Aβ peptide, which can aggregate into fibrils and cause long-term

neuronal injury. The amyloid cascade has led to the identification of potential therapeutic paths: Aβ might be removed by immune-mediated mechanisms induced by vaccination; alternatively the synthesis of Aβ might be blocked by the inhibition of both enzymes involved in the cleavage of Aβ from APP (β-secretase and presenilin-dependent γ-secretase). Finally, the degradation of Aβ might be accelerated by enhancing the activity of Aβ-degrading enzymes. Although the mechanism is still not completely understood, Aβ promotes the deposition of hyperphosphorylated tau, the second pathological marker of the disease.

It is currently believed that Aβ-driven neurotoxicity triggers neurodegeneration, leading to synaptic and neuronal loss. This is indexed by intraneuronal accumulation of abnormally phosphorylated tau, a structural protein constituting microtubules [27]. Beta-amyloid and tau proteins are the main constituents of senile plaques and neurofibrillary tangles, originally described by Alois Alzheimer in the brains, examined under the microscope, of patients with progressive dementia [1]. Pathology studies then showed that, even though AD symptoms generally develop later in life, Aβ and tau accumulate in the brain decades before the clinical onset of the disease [64]. Pre-existing neural reserve and plastic resources of the brain are thought to compensate, for a long time, for the progressive damage caused by Aβ and tau, until a threshold is overcome and symptoms develop (Fig. 2a).

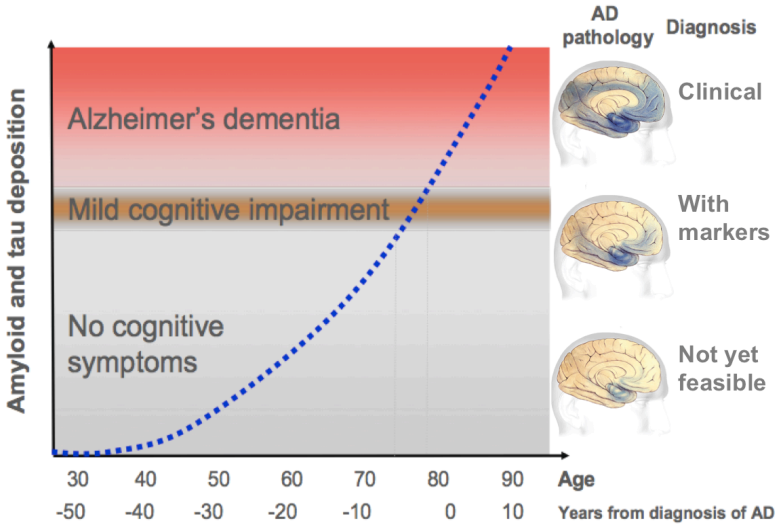


Figure 2A – The natural history of clinical and neurobiological changes. Recent studies have greatly advanced our knowledge of the pathophysiology of Alzheimer's disease (AD), as well as the progression of the neurobiological changes over time. The abnormal accumulation of Aβ and tau in the brain leads to neuronal injury, starting years before the first clinical symptoms appear and proceeding with a

stereotypical pattern of early medial temporal lobe (entorhinal cortex and hippocampus) involvement, followed by progressive neo-cortical damage. The delay in the development of cognitive symptoms suggests that the toxic effects of tau and/or Aβ progressively erode the 'brain reserve' until a clinical threshold is crossed and amnestic symptoms appear. Amnestic mild cognitive impairment (MCI) is the prodromal phase of AD and is characterized by non-disabling memory symptoms. MCI is followed by more widespread cognitive deficits in multiple domains and disability (i.e. lack of self-sufficiency in one or more activities of daily living), when the traditional diagnostic criteria for AD are fulfilled. The case of a person whose brain starts accumulating tau and Aβ around age 40, experiences MCI at age 74, and is diagnosed with AD at age 78 is depicted. The clinical course of the disease lasts only four years, but the neurobiological course lasts almost 40 years.

These acquisitions have allowed the development of new drugs that interfere with Aβ and tau metabolism and accumulation and might halt, or even reverse, the brain damage (Fig. 2b).

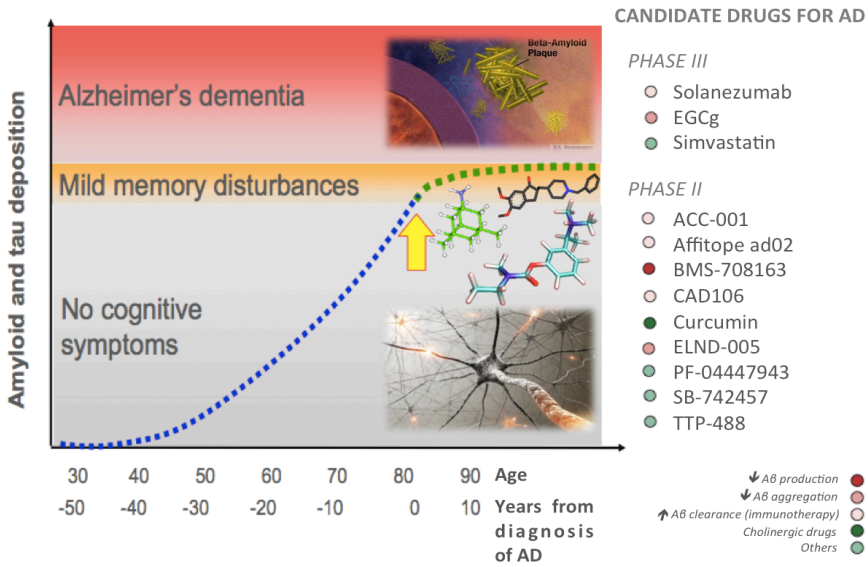


Figure 2B – Tomorrow's therapeutic strategy. The natural history of the neurobiological changes in AD suggests that drugs that might potentially delay the progression of cognitive deterioration should be administered as soon as possible in the course of the disease. The earliest time when it is now possible to recognize the disease is at the stage of MCI. An effective disease-modifying drug administered at this stage might keep the patient in the MCI stage, which is associated with a reasonably good quality of life.

A number of molecules, used in cholinergic, serotonergic, histaminergic, anti-amyloid and tau-related therapies, are currently in phase II and III clinical trials (efficacy studies in patients), and many more are at the preclinical development phase [47, 68, 56].

The development and potential availability of drugs is generating great hopes and expectations, but at least two major hurdles need to be overcome. First, to be maximally effective anti-amyloid and anti-tau drugs need to be prescribed early in the course of the disease. Indeed, the failure of many anti-amyloid drugs, such as: tramiprosate, tarenflurbil, AN1792, and many others, has been attributed, among other things, to the fact that patients were treated in the overt stages of the disease. Second, researchers working on AD will need to find the right set of meaningful surrogate outcomes, or biological biomarkers, sensitive to disease progression which can be effectively deployed to test drug efficacy in a clinical trial setting. These markers can greatly enhance power, allowing up to 10-fold decreases in sample sizes, and thereby making it possible to test a much larger number of drugs and increasing the chances of finding one that is really effective. This approach has already been proven successful in the case of antiretroviral drugs for acquired immunodeficiency syndrome (AIDS), antihypertensive drugs for stroke, and statins for atherosclerosis, whose success was largely due to the availability of relevant biomarkers (blood CD4+ white cell count, blood pressure values, and serum cholesterol levels respectively).

With the advent of the "omics" technologies (i.e., genomics, transcriptomics, proteomics, metabolomics, and connectomics) we entered a new era of biomedical sciences and biomarker discovery. Two-thirds of the approximately 30,000 genes in the human genome are related to brain function, and up to half of the variance in age-related changes in cognition, brain volume and neuronal function appears to be genetically determined. The Ensembl project (the EBI EMBL genome browser sequencing project – http://www.ensembl.org; [17]) has produced a genome database for human and other eukaryotic species, freely available online. In addition, our knowledge of and ability to analyze the transcriptome, proteome and metabolome are now advancing at the same rapid pace that characterized the genomic phase. Besides the specific and complex challenge of identifying and characterizing proteins relevant to pathological AD brains, we need to consider that the overall human proteome is currently estimated to be made up of over 1 million proteins, a staggering amount resulting from: i) single-nucleotide polymorphisms, which cause two-thirds of all the human genes to generate alternative isoforms [50]; and ii) alternative splicing mechanisms of a single gene, which can lead to the codification of different proteins. Since the discovery of these mechanisms, there have been many attempts to catalog the whole human proteome, and in this context, a

special mention must be made of HUPO – Human Proteome Organisation – an international institution fostering proteomic initiatives geared at furthering understanding of human disease (http://www.hupo.org). To date, only few studies have correlated human neuroimaging findings with genomic, transcriptomic, proteomic or metabolomic findings, but such correlations are expected soon.

Proton magnetic resonance spectroscopy (MRS) represents the link with metabolomics, including lipid disorders influencing AD [3]. In patients at risk of AD, MRS can provide a window onto the biochemical changes associated with the loss of neuronal integrity before cognitive impairment arises [28].

"Imaging genetics" is a relatively new branch of neuroimaging which is gaining pace at an unprecedented rate [55]. This methodology exploits an endophenotypes approach in order to identify genes responsible for different cognitive phenotypes. For example, thousands microarrays containing the genetic markers of people with and without good memory, previously assessed through MRI technology, can be compared simultaneously to identify which genes differ and are linked to poor memory performance. Sequencing an entire genome is currently very expensive, but the National Institutes of Health (NIH) hopes that the total cost can be reduced to $1,000 per genome over the next five years.

Other neuroimaging techniques, including PET, MRI, MRS, and functional MRI (fMRI), allow us to investigate the biological macro effects of genetic alterations.

Substantial advances in molecular imaging led to the recent development of PET ligands to track various receptors, neurotransmitters, and proteins, such as Aβ (Pittsburgh compound B, florbetapir, flutemetamol, florbetaben), tau, and acetylcholine [5, 42, 41, 62, 10]. Given that cholinergic deficits as well as amyloid and tau deposition are characteristic of AD, these new ligands should refine our understanding of normal and pathological aging.

A final remark should be made about the administration of neuropsychological tests, the most traditional and pervasive approach to describing and characterizing the stages of AD disease. These tests are a useful means of summarizing, in a single final measurement, all the complex interactive processes described above.

As we have seen, neuroscientists can now leverage on different instrumental techniques to draw up an overall picture of the disease and we believe that all these instruments must now be played together like the strings of a single guitar. Such a multimodal

approach will be instrumental to the success of ambitious scientific initiatives with high societal impact, such as PAD 2020, the Campaign to Prevent Alzheimer's Disease by 2020 (PAD 2020, http://pad2020.org/).

MODELING DYNAMIC CHANGES OF MULTIPLE BIOMARKERS OVER TIME

Over the past two decades, the pathological modification pathway of biological markers over the whole course of AD has been defined. Cross-sectional studies [29, 37, 25] have demonstrated that by the time a patient shows memory deterioration, markers of Aβ and tau deposition in the brain have changed, as have imaging markers of amyloid deposition detectable through [F]AV45 or [C]PIB PET, metabolic markers on glucose ^{18}FDG PET, atrophic markers of neurodegeneration on MRI, markers of axonal and myelin integrity on diffusion tensor imaging (DTI), and markers of neuronal and synaptic function on fMRI (Fig. 3).

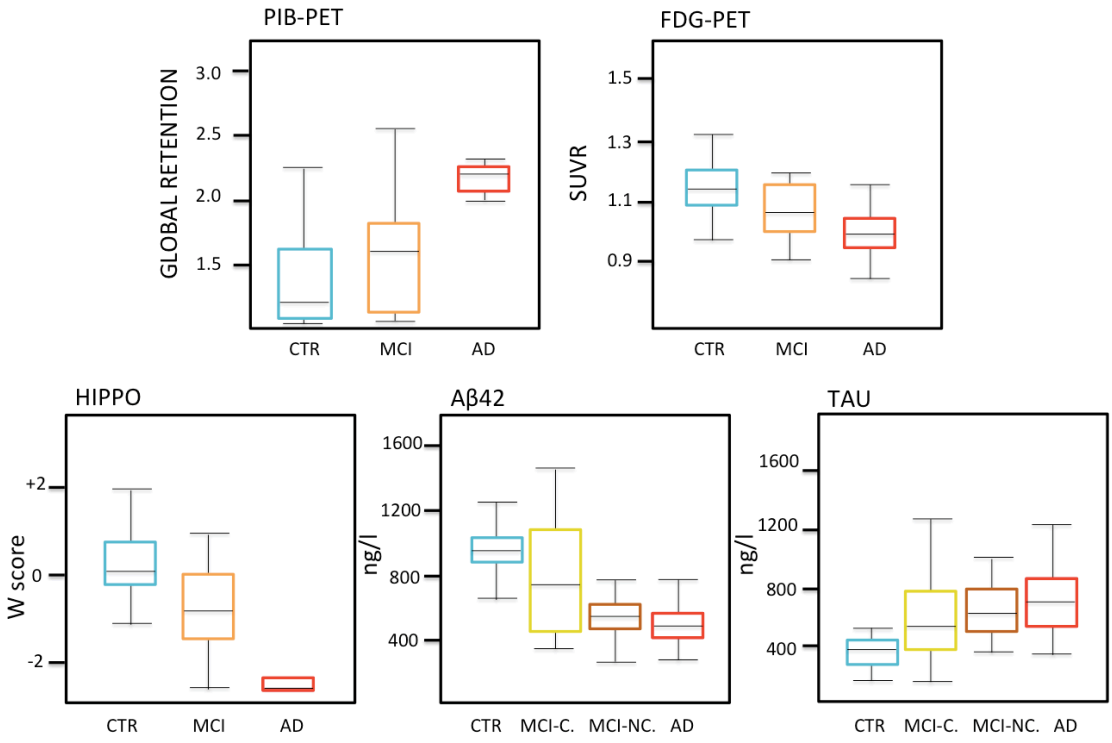


Figure 3 – Markers of Alzheimer's disease. Cross-sectional studies have shown that the pattern of marker abnormality in NC, MCI and AD (see below) is one where controls and AD patients are at the opposite ends (low or high), and MCI patients lie somewhere in the middle. However, the difference between the biomarker level of a patient group and a group of cognitively healthy persons cannot be taken as a point of reference to identify the earliest time of change, since genetic and environmental confounders affect the trajectory of biomarkers. Alzheimer's pathological modifications occur gradually and the dynamics of biomarker changes over time is complex and often non-linear. Acronyms: PIB=Pittsburgh compound B (^{11}C); PET=positron emission tomography; FDG=fluorodeoxyglucose (^{18}F); Hippo=hippocampal volume; Aβ42=beta-amyloid protein ending at amino acid 42; CN=Controls, MCI=mild cognitive

impairment, MCI-C=subjects that convert from MCI state to probable AD stage; MCI-NC=subjects that do not convert from MCI state to probable AD stage; AD=Alzheimer's disease; SUVR=Standardized uptake value ratio; W score=this is the value from a standard normal distribution corresponding to the observed percentile.

All these biomarkers detect unimodal and single-scale changes, and have poor temporal consistency. Moreover, the percentage difference versus a group of matched cognitively healthy persons could not be taken as a reliable indicator of the earliest phase of change over time due to the different metrics and precision of measurements [53].

A great incremental advance in the conceptualization of AD has come from serial studies, initially carried out in Europe on mutation carriers coming from families with AD [18] and more recently from large serial studies, in particular the Alzheimer's Disease Neuroimaging Initiative (i.e.: ADNI-1, -GO and -2) and a related initiative in Australia (Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing – AIBL or "Australian ADNI"). These studies have collected biomarkers serially (generally every 6 to 12 months) in persons with variable degrees of cognitive deterioration, ranging from none, to mild cognitive impairment, through to dementia, allowing neuroscientists to outline the dynamics of the change in biomarkers over time. As an illustrative instance, figure 4 shows that carriers of fully penetrant mutations (i.e., *PSEN1*, *PSEN2* and *APP*), who will inexorably develop AD, exhibit a deviation from the normal trajectory of hippocampal shrinkage as early as 5.5 years before the diagnosis of dementia (the hippocampus is the brain structure where memories are consolidated and where, in AD, tau pathology and neurodegeneration are particularly severe), while a global indicator of whole-brain shrinkage can be detected no earlier than one year before diagnosis. Moreover, when hippocampal shrinkage emerges, atrophy accelerates at a rate of about 0.3% per year [59], showing that the evolution of this biomarker is not linear.

Studies with a similar design, investigating the changes in Aβ42 and tau in CSF, [^{11}C]-PiB amyloid tracer uptake, and [^{18}F]-glucose uptake, have outlined a theoretical scenario wherein some markers change earlier and reach a plateau ahead of others, which instead change later and reach a delayed plateau. In both cases, it is clear that biomarkers follow a non-linear, sigmoid curve [34 - 35].

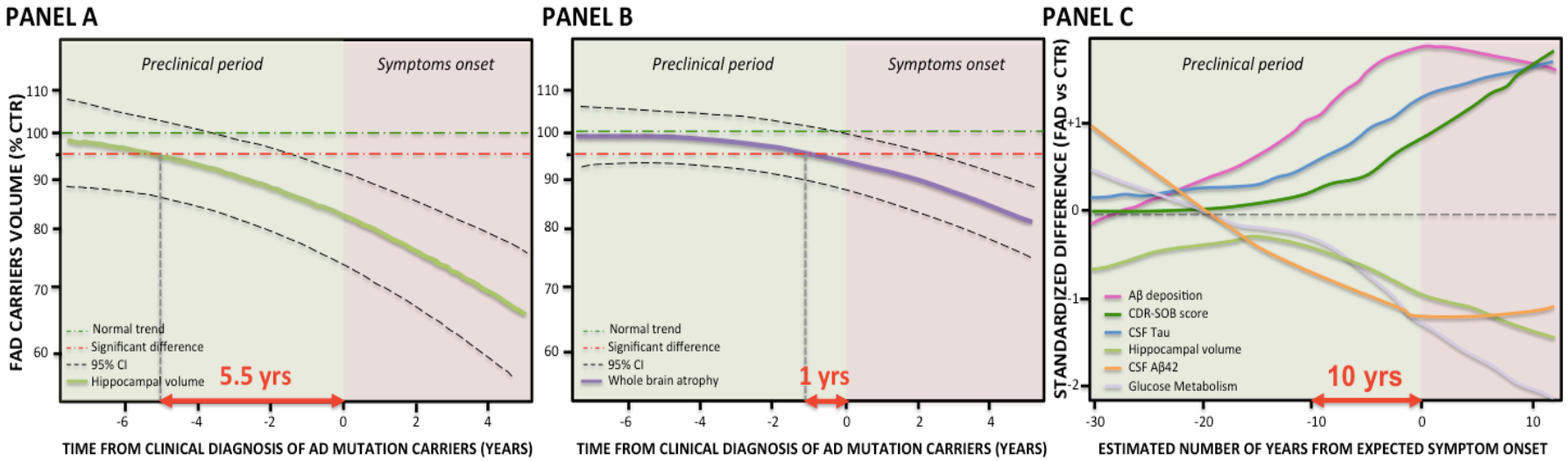


Figure 4 – Dynamics of the change in AD biomarkers over time. Panel A: Ridha et al. (2006) [59] report that persons carrying mutations for FAD provide evidence of acceleration of brain atrophy with disproportionate hippocampal involvement preceding clinical diagnosis. As patients moved from the pre-symptomatic to MCI and AD stages, the mean of total hippocampal and whole-brain volumes decreased. The estimated difference of hippocampal atrophy between autosomal mutation carriers and control groups becomes significant 5.5 years before clinical diagnosis of AD, while the difference in whole-brain atrophy only around 1 year before diagnosis (Panel B). Moreover, once hippocampal and whole-brain shrinkage has appeared, atrophy rates accelerate, showing a non-linear change over time. Panel C: The Dominantly Inherited Alzheimer's Network (DIAN) and studies of Colombian kindred carriers of a *PSEN1* mutation support the idea of a protracted preclinical period (10 years or more) during which biomarkers become abnormal sequentially while people remain clinically asymptomatic. Additionally, the DIAN results suggest that CSF Aβ42 might become abnormal before amyloid PET, with CSF Aβ42 initially starting at abnormally high concentrations followed by a progressive decline. DIAN results also suggest that tau becomes abnormal before FDG PET and that FDG PET and MRI become abnormal very close in time. Standardized difference in Panel C is derived from the mutation carrier group and the non-carrier group showing the non-linear changes over time. Abbreviations: Yrs=years; FAD=Familial Alzheimer's disease; CTR=Controls; CSF=cerebrospinal fluid; Aβ42=beta-amyloid protein ending at amino acid 42; CDR-SOB=Clinical Dementia Rating Scale Sum of Boxes.

Depending on the degree of abnormality and the slope of change, different biomarkers at different times can be used for diagnosis or monitoring progression over the disease course. The recently conceptualized scenario suggests that the first pathological event consists of the brain changes taking place at the molecular level (toxic amyloid deposition), which lead to the destruction of synaptic functions and axonal integrity; these are followed by neuronal loss, gray and white matter atrophy and, finally, clinical cognitive decline [67].

Modeling the changes in biomarkers over time is useful for many reasons. It allows us to predict which biomarker or combinations of biomarkers are more sensitive to the disease state, with practical implications for the diagnosis, and which to the disease progression, with the possibility of better understanding the efficacy of disease-modifying drugs. Figure 4 shows that hippocampal atrophy might be of poor diagnostic value at the earliest stages of the disease, but might be an accurate and valuable marker of progression later on in its course. On the contrary, amyloid burden might be a good diagnostic marker in the earliest stages of the disease, but might be poorly useful as a marker of disease progression. When modeled biomarkers are used as the basis of a prediction, they allow the formulation of a robust pathogenic hypothesis. For example, according to the modeled rates of atrophy, the medial temporal cortex changes substantially in the early stages of the disease, while rates of atrophy in the frontal cortex are either flat or occur at a later stage. What this shows is that neurodegeneration starts earlier in the medial temporal cortex and only later spreads to the frontal cortex.

In the next section, we will conceptualize what we have called the hypermodel of AD pathology. The main point on which the following dissertation hinges is the capability of the brain simulator to estimate the multimodal and multiscale pathological changes over time with the highest possible consistency [63]. The hypermodel will transform the diagnosis and treatment of brain diseases, providing insights into the organizational complexity of this convoluted organ.

THE BRAIN HYPERMODEL: A MULTISCALE AND MULTIMODAL DYNAMIC SIMULATOR

Definition of the brain hypermodel

Modern neuroscience has afforded deep insights into every level of brain organization – from genes to cognition. Neuroscience today is faced with the compelling need to fit the different levels together, exploiting advanced theoretical models, such as the hypermodel of the brain, in order to capture, through solid mathematical rules, all the deep mechanics of the brain. Thus, our knowledge of AD can be improved [15]. This multiscale challenge is made even more complicated by the temporal and spatial scales at play, which range over nine-ten orders of magnitude (Table 1).

	Scale	CTR	MCI	AD
Genomics (SNP analyses)	nm (10^{-9})	Carriers of protective alleles	Carriers of protective alleles	Carrier of risky alleles
Transcriptomics (Alternative splicing & promoter regions analyses)	nm (10^{-9})	mRNA transcripts from: *BACE1, APOE, COMT, GRM3, serotonergic genes* and *APP* protective gene	mRNA transcripts from: *BACE1, APOE, APP, PSEN, COMT, BDNF, ACE, NOS, GRM3, KIBRA, IGF, SOD1, PLXNB3, MTHFR* genes	mRNA transcripts from: *BACE1, APOE, APP, PSEN1-2, COMT, BDNF, ACE, NOS, GRM3, KIBRA, IGF, SOD1, DAPK1, DISC1, MTHFR, BCHE, CADPS2* genes
Proteomics (sequence and functions analyses)	10 nm (10^{-8})	Systematic characterization of: homocysteine, C-reactive protein, α2-macroglobulin, complement factor H isoforms	Systematic characterization of: homocysteine, C-reactive protein, α2-macroglobulin, fibrinogen, GSK-3, complement factor H isoforms	Systematic characterization of: Homocysteine, C-reactive protein, α2-Macroglobulin, Oxysterols, F2-isoprostanes, DKK-3, complement factor H isoforms
CSF / Blood markers	10 nm (10^{-8})	T-TAU → P-TAU → Aβ42 → Inflammation →	T-TAU ↑ P-TAU ↗ Aβ42 ↓ Inflammation →	T-TAU ↑ P-TAU ↑ Aβ42 ↓ Inflammation ↑
Metabonomics (MRS)	10 µm (10^{-5})	NAA/Cr ratio ↑ mi/Cr ↘ Cho/Cr ↘	NAA/Cr ratio ↘ mi/Cr → Cho/Cr →	NAA/Cr ratio ↓ mi/Cr ↗ Cho/Cr ↗
Synaptic functions (FDG PET)	1/10 mm (10^{-4})	Glucose consumption unimpaired →	Glucose consumption very mild impaired ↘	Regional decline of glucose consumption ↓
Functional (rsfMRI), Structural (T13D) and Connectivity (DTI)	mm (10^{-3})	Functional → Structural → Connectivity →	Functional ↘ Structural ↘ Connectivity →	Functional ↓ Structural ↓ Connectivity ↓
Cognitive / Behavioral tests	meter (10^{1})	- CDR → - MMSE → - Risk of cogn.decline ↓ - Intact behavior	- CDR ↗ - MMSE ↘ - Risk of cogn.decline ↑ - Minimally impaired behavior	- CDR ↑ - MMSE ↓ - Dementia developed - Clinically impaired behavior

Table 1 – The brain hypermodel relies on multiple biomarkers describing the dynamics of the brain in the different stages of the disease. Abbreviations: AD=Alzheimer's disease; CTR=healthy elderly control; MCI=mild cognitive impairment; CSF=cerebrospinal fluid; mRNA=messenger ribonucleic acid; SNP= single nucleotide polymorphism; *ACE*=gene encoding for angiotensin-converting enzyme; *APOE*=gene encoding for apolipoprotein E; *APP*=gene encoding for amyloid precursor protein; *BACE1*=gene encoding for beta-secretase1;

BCHE=gene involved in butyrylcholinesterase synthesis; *BDNF*=gene encoding for brain-derived neurotrophic factor; *CADPS2*= gene encoding for calcium-binding protein involved in exocytosis of vesicles filled with neurotransmitters and neuropeptides; *COMT*=gene encoding for catechol-O-methyl transferase; *DAPK1*=gene encoding for death-associated protein kinase 1; *DISC1*=gene implicated in thought and working memory; *GRM3*=gene encoding for metabotropic glutamate receptor; *IGF*=gene encoding for insulin growth factor; *KIBRA*=gene involved in hippocampal activation; *MTHFR*=gene encoding for methyl-tetrahydrofolate reductase; *NOS*=gene encoding for nitric oxide synthase; *PLXNB3*=this gene is a member of the plexin family playing a role in axon guidance; *PSEN1-2*= genes encoding for presenilin1 and 2; *SOD1*=gene encoding for superoxide dismutase 1; T-TAU=total tau; P-TAU: phosphorylated tau; AB42=Abeta 1-42 protein in CSF. ↑=increased ↗=slightly increased ↓=decreased ↘=slightly decreased →=stable. MRS=magnetic resonance spectroscopy. Cho=choline; Cr=creatine; mi=myo-inositol; NAA=N-acetyl aspartate; FDG-PET= ^{18}F-fluorodeoxyglucose positron emission tomography. MRI=magnetic resonance imaging; rsfMRI=resting state functional MRI; DTI=diffusion tensor imaging; CDR=Clinical Dementia Rating scale; MMSE: Mini Mental State Examination.

The hyper brain is a mathematical model, based on the Bayesian inference, useful to describe brain activity ranging from the low-level mechanisms up to the large-scale biological processes [23 - 24]. To understand the Bayesian brain it is necessary to understand the structure and connectivity hierarchically linking the variables considered in the analysis. The general assumption is that genes influence cognition and behavior. Therefore, the first step involves regulation and transcription of genes in many proteins. These proteins eventually influence cell processes and functions through enzymatic reactions. Neurons, forming neural networks, work together in a complex pattern of stimulation and inhibition, along with other interactions, to produce a given cognitive behavior. Taking this into consideration, neuroimaging techniques, including PET, MRI, MRS, and fMRI, allow us to examine the biological effects of genetic alterations. Both at genetic and neuroimaging level, hierarchical Bayesian mixture models have started to be proposed [70, 51]. The structure and function of the human brain in these models can be studied at multiple temporal and spatial scales.

BOX 1 should help to familiarize the reader with the basic Bayesian components adopted by the brain hypermodel.

BOX 1 / The brain hypermodel: basic concepts

Hierarchical model: The model is described by several parameters that vary at more than one level. The hierarchical model is suitable in cases of nested data (e.g., omics, imaging, clinical, patient's neuropsychological data, etc.). In this hierarchical analysis, the estimated elements come from subjects randomly selected from a larger population.

Bayesian model: A typical Bayesian model concerns the probabilistic relationships between diseases and symptoms. Given a set of symptoms, a Bayesian network can be used to compute the probabilities of the presence of a certain disease. Brain hypermodels typically rely on a Bayesian network, a probabilistic model expressed via a graph (Fig. 5): here, every node of the graph represents random variables (e.g., observable quantities, unknown parameters or hypotheses), edges represent dependencies and those nodes which are not connected represent independent variables. Each node is associated with a probability function that takes a particular set of values from the node's parent as input and gives the probability of the variable represented by the node as output. The Bayesian model relies on additional concepts; i.e.: (A) *Prior probability*: this is the probability distribution that confers the uncertainty on an uncertain quantity defined as "p" (e.g.: suppose "p" is the number of voxels that will be activated for a specific task in an fMRI experiment) before the data are taken into account (in this case, the results of the experiment obtained via independent component analysis); (B) *Likelihood*: this is synonymous of probability, albeit with some differences. Probability is used when describing a function of the outcome given a fixed parameter value, and it can be described as follows: "if a transcript of a messenger ribonucleic acid (mRNA) is expressed 100 times and this mRNA is not affected by errors from the RNA polymerase, what is its probability of expressing a fully functional protein?". The term likelihood, instead, is used when describing a function of a parameter given an outcome. For example: "if an mRNA is translated 100 times and it encoded for an active protein 100 times, what is the likelihood of the mRNA being unaffected by errors?"; (C) *Posterior probability*: this measures the likelihood that an event will occur given that a related event has already occurred. An example can be given by calculating the probability of case of MCI converting to AD, given that the level of Aβ42 in the CSF has decreased. Let A be the event that MCI converts to AD, and the probability that MCI will convert is 75% ($P(A) = 0.75$). Let B be the event that the level of Aβ42 decreases, with a probability of 80% ($P(B) = 0.80$). Finally, let the likelihood that Aβ42 will decrease, given that MCI converts to AD, be 99% ($P(B|A) = 0.99$). The probability that MCI will convert to AD given that Aβ42 decreases can be determined by plugging these values into the Bayes' Theorem, giving $P(A|B) = {}^{0.99*0.75}/_{0.80} = 0.92$. This means that in this hypothetical situation if the CSF Aβ42 level is decreasing, MCI has a 92% chance of converting.

Multivariate analysis (MVA): This statistical technique is based on observation and analysis of more than one outcome variable at a time. The technique is used to perform studies across

multiple dimensions while taking into account the effects of all variables on the responses of interest.

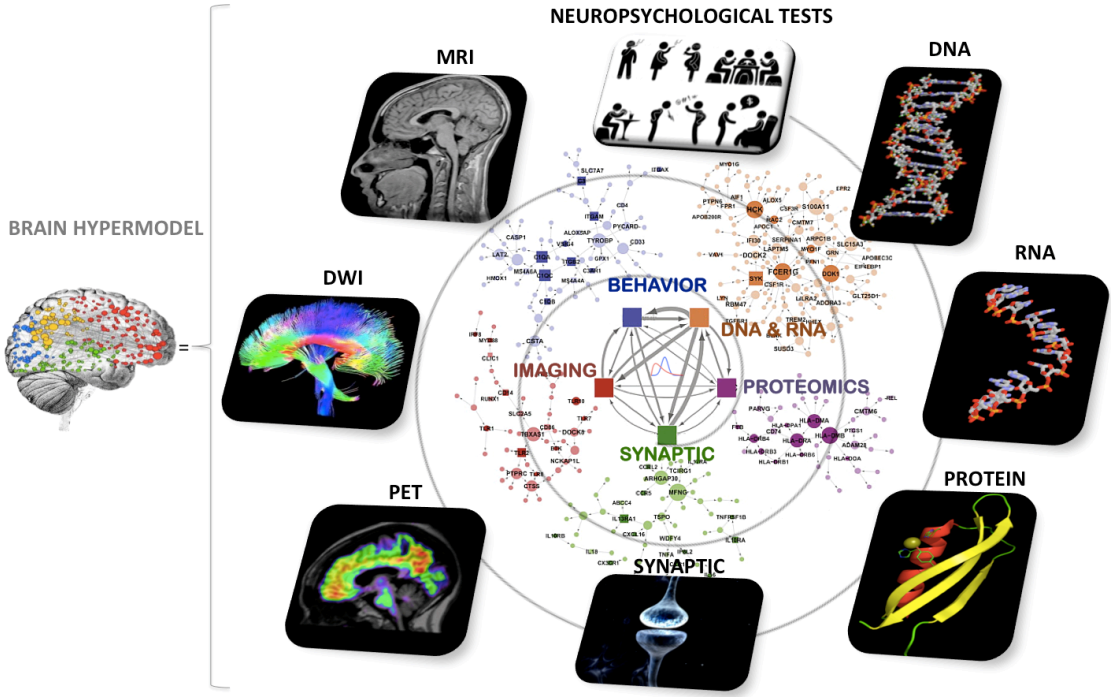


Figure 5 – Definition of the brain hypermodel. This is represented as a set of mathematical relations expressed in terms of random variables and associated probability distributions with the aim of describing the observations of brain atrophy merging with different levels of information. These levels are dynamically linked through a directed acyclic graph (DAG). The set of equations describes the longitudinal mutual biomarker variations as well as their temporal and spatial interactions from the lowest to the highest scale.

To develop the brain hypermodel of AD based on genetic, clinical, imaging and behavioral data, a large number of postprocessing tools are required (Fig. 6) in order to generate inputs that feed the Bayesian network. BOX 2 will help the reader to understand the programs and ITC tools used by the brain hypermodel.

BOX 2 / ITC tools needed for the brain hypermodel

"OMICS" tools: Being concerned with strands of nucleotides [i.e., deoxyribonucleic acid (DNA) or ribonucleic acid (RNA)] and chains of amino acids (i.e., oligomers or proteins), the hypermodel should handle inputs coming from sequence analysis tools (e.g.: Genscan, ExPASY, ORF-finder), sequence alignment programs (e.g., BLAST, PipeAlign,

PROMALS3d), and monitoring protein expression algorithms (e.g., CountCodon, Molecular Toolkit, Promoser).

Imaging tools: The model needs specific algorithms to properly register different brain image modalities of different subjects at different time points into the same space. Modeling the course of brain changes in neurodegenerative disorders requires spatial consistency at multiple spatial scales. The Boundary Shift Integral (BSI) is one of the tools available to segment and register brain scans at multiple time points [44]. Other single image analysis tools for AD are: the Multi-Atlas Propagation Segmentation (MAPS) for the hippocampus, which combines atlas-based segmentation and multi-feature pattern recognition [43]; the PIB model uptake [61]; the multi-atlas based anatomical segmentation tool [71, 46]; DEMONS, a deformation-based patient normalization and follow-up method [69]; and the 4D longitudinal brain atrophy simulation tools to predict brain atrophy. New insights into microstructural changes of the white matter should be assessed through deformation-based morphometry (DBM) [73]. All these algorithms can play an important role in the definition of the hypermodel. However, the aforementioned tools are needed to pre-process multimodal data that must then be analyzed through advanced imaging libraries and tools to assess the final biomarkers. Therefore, additional algorithms to be adapted and plugged in the model are: FSL [72], FreeSurfer [16, 6], Civet [40], BrainCSI for MRS, Voxel/Bayesian based morphometry [31] and BrainVISA [9].

Neuropsychological tools: the integration of the neuropsychological data can be done by interfacing computerized assessment tools such as CANTAB [14].

The level of complexity implicit in connecting and integrating all these single-modality tools into a consistent multimodal framework (the hypermodel) is such as to drive the evolution of the research framework from the traditional models, such as a generalized linear model (GLM), to hierarchical (mixed effects) models.

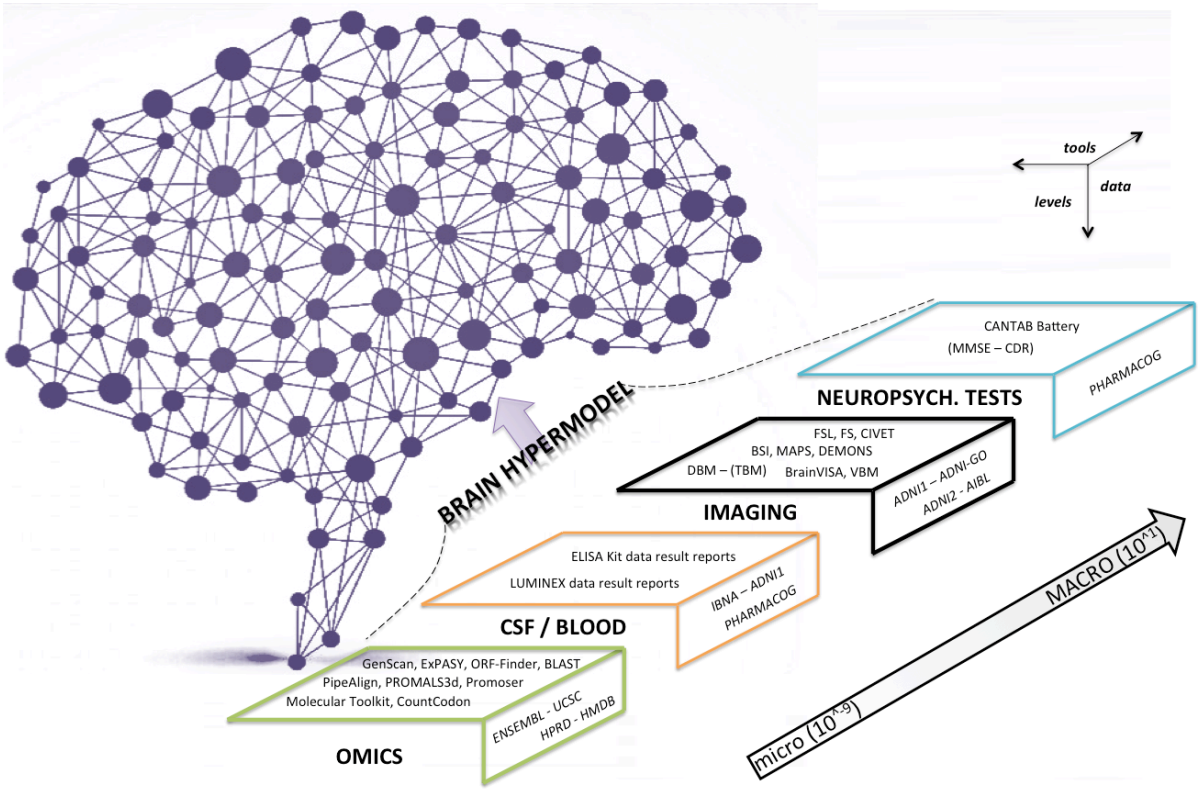


Figure 6 – Tools and data exploited by the brain hypermodel. The scheme represents algorithms and datasets structured by different levels of depth and scale. The acronym and URL for each component are explained below. Genscan = program to identify the gene structure from the DNA strand (http://genes.mit.edu/GENSCANinfo.html); ExPASy = Web-portal to obtain user access to proteomics, genomics, phylogeny, systems biology, population genetics and transcriptomics (http://www.expasy.org/); ORF-Finder = Open Reading Frame finder in RNA coding strand (http://www.ncbi.nlm.nih.gov/projects/gorf/); BLAST = algorithm able to find regions of similarity between biological sequences (http://blast.ncbi.nlm.nih.gov/); PipeAlign = toolkit for protein family analysis (http://bips.u-strasbg.fr/PipeAlign/); PROMALS3d = multiple protein sequence and structure alignment tool (http://prodata.swmed.edu/promals3d/promals3d.php); Promoser = tool for transcription regulation analysis (http://cagt.bu.edu/page/Promoser_about); Molecular Toolkit = tool for manipulation of nucleic acids and protein (http://www.vivo.colostate.edu/molkit/); CountCodon = on-line tool to count codons in mRNA (http://www.kazusa.or.jp/codon/countcodon.html); FSL = complete library for the analysis of fMRI, MRI, DTI data (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/); FS = FreeSurfer, a set of automated tools for reconstruction of the brain's cortical surface from structural MRI data (https://surfer.nmr.mgh.harvard.edu/); CIVET = tool for the segmentation of the cerebral cortex (http://cbrain.mcgill.ca/); BSI = Boundary Shift Integral (http://idealab.ucdavis.edu/software/bbsi.php); MAPS = multiple-atlas propagation and segmentation tool; DEMONS = diffeomorphic registration algorithm (http://www.insight-journal.org/browse/publication/154); DBM – (TBM) = deformation based morphometry; BrainVISA = complete set of tools and libraries to process brain image data (http://brainvisa.info/); VBM = voxel-based morphometry (http://www.fil.ion.ucl.ac.uk/spm/); CANTAB = Cambridge Neuropsychological Test Automated Battery; ENSEMBL = genome browser sequencing project (www.ensembl.org); EBI EMBL = The European Bioinformatics Institute (http://www.ebi.ac.uk); UCSC

= University of California Santa Cruz genome and transcriptome browser (http://genome.ucsc.edu/); HPRD = Human Protein Reference Database (http://www.hprd.org/); HMDB = Human metabolome database (http://www.hmdb.ca/); IBNA = Italian Brain Normative Archive; ADNI (1-GO-2) = Alzheimer's Disease Neuroimaging Initiative (http://adni.loni.ucla.edu/); AIBL = Australian ADNI (http://www.aibl.csiro.au/); PHARMACOG = European ADNI dataset (http://www.imi.europa.eu/content/pharma-cog). OMICS = of or pertaining to related measurements or data from fields such as genomics, proteomics, transcriptomics, metabolomics; CSF = cerebrospinal fluid; NEUROPSYCH. TESTS = Neuropsychological tests.

Brain hypermodel axioms

The brain hypermodel needs to be based on specific fundamental principles:

- Current observation depends on past observation.
- The distribution (i.e.: prior, likelihood and posterior) of every biomarker has to be derived according to Jack (2013) [34].
- Multi-level descriptions of the brain space, ranging from genes to proteins, from microcircuits to voxels, from small tissues to global regions of interest, must come from large serial datasets, such as: ENSEMBL – (www.ensembl.org); UCSC – University of California Santa Cruz genome and transcriptome browser (http://genome.ucsc.edu/); HPRD – Human Protein Reference Database (http://www.hprd.org/); HMDB – Human metaboloma database (http://www.hmdb.ca/); and ADNI – Alzheimer's Disease Neuroimaging Initiative (which comprises huge number of image modalities e.g.: [^{11}C]PIB PET, [^{18}F]- AV45-PET, [^{18}F]FDG PET, resting fMRI, DTI and structural MRI) (http://adni.loni.ucla.edu/). Additional datasets might need to be added for further refinement of the brain hypermodel.

Brain hypermodel statistical pillars

The most obvious approach to the modeling of dynamic, multimodal and longitudinal measurements is through hierarchical (or random effect) models, as used in many recent publications [59]. A hierarchical model offers the advantage of modeling the spatial dependence of variables at neighboring locations using multilevel descriptions of the space at scales ranging from local (nucleotide) to global (voxel regions of interest or even lobes).

The use of hierarchical models, with empirical Bayes estimation, in the field of neuroimaging was initially proposed by Friston et al. in the context of fMRI data analysis, as a way of overcoming some constraints and limitations of the classical statistical parametric mapping approach. In the analysis conducted by Friston et al., this technique made it possible, in contrast

with the classical statistic approach, to move from a local (e.g. at voxel level) to a global (at whole-brain level) estimation, with the tangible benefit of increasing the power in the detection of statistically significant results.

At all spatial scales, however, a statistical issue may arise that needs to be taken into account, namely, the false positive detection rate due to multiple comparisons. However, techniques for a posteriori correction of results are available, based on both parametric (i.e., Bonferroni, false discovery rate, family-wise error) [32, 22, 26] and non-parametric assumptions (i.e., bootstrap, permutation tests) [52].

The hierarchical formulation might benefit from the so-called multivariate exchangeability assumption. This approach allows missing data to be substituted in order to promptly estimate the subject's parameters. What this means is that, if necessary, the individual estimated subject parameter can gain in consistency thanks to an increased weight, assumed from the estimation gathered from the entire population, and thus move from a poor subject estimate to a wider and well-defined population perspective. This is one of the key advantages offered by hierarchical Bayesian modeling as opposed to the classical regression approach.

Flexibility and added value of the brain hypermodel

The hypermodel might be considered a high-order marker of disease progression that could be highly representative of all the data. While many parameters will provide direct information about the progression of the disease, others might give "clues" as to the right direction to explore, and provide new insights for a better explanation of data. A statistical analysis of the hypermodel, considering the correlation and redundancy between the variables, could identify significant spatial patterns and time trends.

Unfortunately, the brain hypermodel can still be hampered by a very large number of variables. In the practical clinical setting (e.g. in clinical trials and for early/differential diagnosis), a reduced or simplified model might be used for a more predictive and individualized healthcare. Examples of this model have been proposed showing incredibly high diagnostic and prognostic power [65]. Predict-AD (http://www.predictad.eu/), a recently EU-funded research project, has developed and adopted objective and efficient methods to enable earlier diagnosis of AD through a holistic view of patients which combines information from several sources, such as blood samples, imaging and clinical tests [2].

Additionally, the hypermodel could estimate the deviation from the "experienced curve" of neurodegeneration during a clinical trial with a disease-modifying agent, a deviation

translatable into a measure of treatment efficacy. Modeling the adverse effects of a treatment will allow researchers to assess the safety of new drugs, which is a critical step on their route to market and an aspect that in the past has proved to be a common cause of expensive failures [45, 57]. There is evidence that side effects of new AD drugs might include micro-bleeds and inflammation [11]. Even if subjective assessment of radiological images can be used to detect these kinds of side effects, these measures are relatively crude and lack quantification. In this regard, the hypermodel might make a significant contribution to imaging safety in the context of biomarker quantifications.

Finally, the brain hypermodel might help to overcome current limitations in early detection and clinical management of dementia due to lack of sensitive and specific biomarkers for classification and prediction. Specifically, the hypermodel could locate a given patient, studied at one point in time, on the appropriate trajectory (e.g. healthy or AD), and from there predict past and future points (e.g. five years before symptoms, one year after symptoms, etc.) according to the specific pattern of his/her disease marker evolutions.

THE COMPUTATIONAL ENGINE

To overcome the high computational needs required by a multimodal and multiscale brain model, we describe here the most well-equipped e-infrastructures available worldwide that can host the Bayesian model and its processing tools, to perform ad hoc brain hyper simulations on real data. First, what is an e-infrastructure? An e-infrastructure offers neuroscientists advanced image analysis algorithms, powerful resources, 3D visualization tools, quality control services as well as statistical tools, a fertile ground for brain hypermodels. An e-infrastructure allows neuroimaging experiments to be conducted using dedicated computational resources such as: grids, high-performance computing (HPC) systems, and clouds. The remarkable growth, accessibility and availability of imaging and non-imaging data from people affected by neurodegenerative conditions have recently fostered the development of many of these computational e-infrastructures [20].

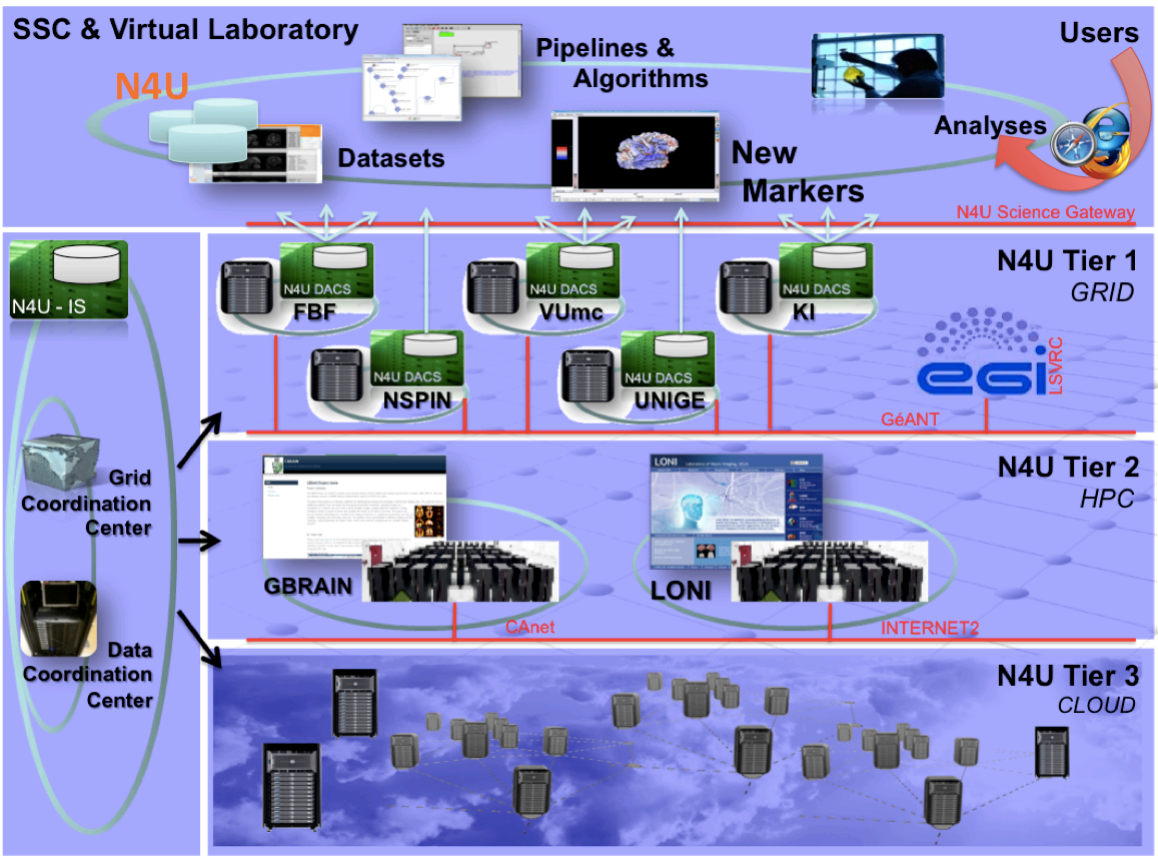


Figure 7 - The N4U infrastructure. The infrastructure is composed of different layers. At the top, there is the Specific Support Centre (SSC) and the Virtual Laboratory, accessible via browser. The goal of the SSC is to support users, providing training and assistance such as: designing a scientific experiment, building the brain hypermodel, uploading data, customizing algorithm pipelines, running different kinds of analyses, visualizing

outputs, checking results, and carrying out statistics. The Virtual Laboratory provides the environment for users to conduct their own analyses on specific sets of data. The neuGRID platform is composed of:

(I) An on line web-portal (https://neugrid4you.eu) to provide facilities for users to interact with the neuGRID services. Users can leverage on the Online Help Desk, a one-stop assistance facility with which every neuroscientist can interact in order to learn about neuGRID and how to use it.

(II) A set of data resources. Data can be integrated in N4U upon users' request. In neuGRID, all data are indexed and registered creating a user-friendly atlas. All data, pipelines and experimental results can be browsed and queried.

(III) An analysis work area. Here neuroscientists can define new pipelines or configure existing algorithms to be run against selected datasets. At the end, results can be visualized.

(IV) Access to the quality control and statistical tool environment providing neuroscientists with informative reports on the execution of their pipelines.

The neuGRID virtual laboratory sits on top of a 3-tier distributed computing infrastructure:

(a) Tier 1: this is the real core of the infrastructure. It is composed of a number of sites providing computing resources and integrated services. The sites are located in Italy, France, Sweden, Switzerland and The Netherlands.

(b) Tier 2: this attaches additional public facilities augmenting N4U's capacity (e.g. LONI, CBRAIN, ESFRI).

(c) Tier 3: this adds private cloud computing resources from external providers.

The three different tiers are coordinated by the Grid Coordination Center (GCC) and the Data Coordination Center (DCC). The DCC coordinates the neuGRID data, quality control and analysis procedures. The GCC is in charge of hosting, maintaining and running the grid computing system services. These services are the cornerstones of the platform providing the inner mechanics of the neuGRID grid/cloud job parallelization.

Amongst these, neuGRID (www.neugrid4you.eu) is the leading European e-infrastructure, developed with the aim of overcoming those hurdles that each neuroscientist has to face daily when trying to set up an advanced experiment on computational neuroimaging. Here neuroscientists can find core resources for their analyses. The neuGRID platform offers access to 500 processing cores, 25 terabytes of effective storage and it has established a connection with external computing resources to double its capacity on demand. From a bandwidth point of view, neuGRID leverages on the pan-European research and education network GEANT (www.geant.net). Although originally designed for neuroscientists working on AD, neuGRID has, in a second phase, been expanded to deal with a wider range of brain diseases, such as to the white matter disease and psychiatric diseases. NeuGRID also includes tools useful for clinical use, sensitive to the departure of single cases from a normative reference image database (Fig. 7).

LONI (Laboratory Of Neuro Imaging – http://www.loni.ucla.edu/) focuses on the development of image analysis methods and their application to health issues. The LONI e-infrastructure is the longest-established platform among those available in the field of neuroimaging [13]. It responds to the needs of a wide range of users, offering specific services (data and algorithms) to both neuroscientists and neurobiologists. LONI hosts the ADNI databases, which comprise clinical data and

	neuGRID	LONI	CBRAIN
		HARDWARE AND CONNECTIVITY FEATURES	
Infra topology	Distributed using the GRID and CLOUD paradigm	Centralized using HPC	Distributed using the GRID and HPC paradigm
Storage capacity	25 TB + distributed storage	12 PB (at UCLA/LONI)	200 TB + distributed storage
Core resources	500 CPU cores (plus EGEE external 10,000 cores)	10,000 CPU cores	Over 40,000 CPU cores (plus JUROPA external 50,000 HPC cores)
Network provider	GEANT	Internet2	CANet
Bandwidth	1 GB/s	2*10 Gb/s (load balanced)	10 GB/s
Platform link	www.neugrid4you.eu	http://pipeline.loni.ucla.edu/products-services/pws/	https://portal.cbrain.mcgill.ca/login
		ALGORITHMS & PIPELINES	
Target community	AD, WMD, PSY	AD, PSY	AD
Processing algorithms	Imaging tools for structural (T13D, T2), functional (PET, rsfMRI) and diffusion (DTI) imaging analysis.	Imaging tools for structural (T13D, PD, T2), functional (PET, rsfMRI, ASL) and diffusion (DTI) imaging analysis. Bioinformatic tools for conducting analysis on GWAS, BLAST, sequence alignment tools.	Imaging tools for structural (T13D) and functional (PET) imaging analysis .
Statistical tools	R, Octave	Many different tools covering data classification, linear and non-linear regression, feature selection, and multivariate analysis	R and the Rminc package. Integrated voxel-based statistics and voxel wise-linear models.
DBMS	Cristal	SQL	MySQL
		IMAGE DATASETS	
AD	OASIS; ADNI (through LONI); 1000 Functional Connectomes - INDI; ADHD-200; PharmaCOG	ADNI1; ADNI-GO; ADNI2; AIBL; ABIDE; BRIN; PAD/CRYO	ADDNEUROMED
WMD	LADIS; EDSD		
PSY	FBIRN; ELUDE	ABIDE	NIHPD

Table 2 - Main features of the three e-infrastructures in terms of (I) Image data sets available; (II) Image-processing algorithms, suites and tools available; (III) resources and connectivity. Abbreviations: TB = terabytes; PB = petabytes; EGEE = Enabling Grids for E-Science in Europe. This is a public resource expanding the computational power of the neuGRID platform; LONI = Laboratory of Neuro Imaging; UCLA = University of California, Los Angeles; CPU = central processing unit; GB/s = gigabytes per second; AD = Alzheimer's disease; WMD = white matter disease; PSY = psychiatric disease; T13D = Volumetric sequence weighted in T1; T2 = MRI sequence weighted in T2; PD = proton density-weighted image; PET = positron emission tomography; rsfMRI = resting-state functional MRI; DTI = diffusion tensor imaging; GWAS = genome wide association study; BLAST = basic local alignment search tool; ADNI = Alzheimer's Disease Neuroimaging Initiative; AIBL = The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ABIDE = Autism Brain Imaging Data

Exchange; BRIN = Brain Info; PAD/CRYO = Public Anonymized Dataset/Cryosection; ADDNEUROMED = the AddNeuroMed study; NIHPD = National Institutes of Health Pediatric Database; OASIS = Open Access Series of Imaging Studies; 1000 Functional Connectomes – INDI = 1000 functional connectomes International Neuroimaging Data Sharing Initiative project; LADIS = Leukoaraiosis And DISability; EDSD = European diffusion tensor imaging study in dementia; FBIRN = The Functional Bioinformatics Research Network; ELUDE = Efficient Longitudinal Upload of Depression in the Elderly.

information from genetic scans from older people with AD (400 mild AD), people with mild cognitive impairment (350 early MCI, 400 MCI and 150 late MCI), and healthy elders (350 CTR). The LONI imaging portfolio comprises high-resolution structural MRI (T13D MPRAGE, T2, PD), ^{18}F-FDG PET, amyloid PET (AV45-PET), fMRI and DTI. Algorithms for data analysis are available via the LONI Pipeline graphical interface, a user-friendly workflow management system that makes it possible to automatize the measurement of functional, tractographic and morphometric analyses, to dynamically assess volume and shape features, and to extract and associated cognitive, genetic, clinical, behavioral and imaging biomarkers. LONI provides access to a large, centralized HPC infrastructure – located at the University of California, Los Angeles (UCLA) – for computationally intensive analyses. External researchers are granted access to the LONI HPC resources on the basis of ad hoc scientific agreements.

CBRAIN (http://cbrain.mcgill.ca/) is a network of five Canadian brain imaging research centers, connected to HPC centers in Canada and Europe. The CBRAIN e-infrastructure offers advanced networking, transparent access to computing resources, a wide range of tools as well as web-based results visualization, all thanks to a comprehensive and well organized web portal. CBRAIN is a distributed environment connected through a high-speed wide area network bandwidth.

Table 2 summarizes core features, datasets and tools of the three leading e-infrastructures available in the field of neuroimaging.

NeuGRID is expanding its platform internationally, bridging with the other e-infrastructures, with the ultimate aim of delivering an authentic Virtual Laboratory, integrating the widest range of available analysis services with a specific support center for end users. This will create a virtual space accessible by the user via web no matter where he/she is physically located.

The above facts and figures support the notion that e-infrastructures are today the most advanced and the best equipped platforms to support the deployment and distribution of the hypermodel of the brain. In this way, a neuroscientist would be just a click of his/her fingertips

away from all he/she needs to start a simulation. Along the same lines, the recent EU FET Flagship initiative called The Human Brain Project (HBP: http://www.humanbrainproject.eu/) as well as the American Brain Activity Map project (BAM: http://www.nih.gov/science/brain/) will widely exploit the larger scale of data and the huge power of the resources available through NeuGRID, LONI and CBRAIN to characterize, build and test the respective *in silico* brain models.

CONCLUDING REMARKS

According to the latest EU estimates, the global prevalence of AD is predicted to quadruple to reach 105 million by 2050. To tackle this social emergency and improve AD diagnosis over the next 15 years, clinical practice will need to rely more and more on multimodal methodologies, using an integrated approach based on genetic, biological, imaging methods, as well as neuropsychological and cognitive tests. Indeed, modern neurobiology and neuroscience have gained deep insights into every level of brain organization, and this will help us to move closer to the real cause of the disease rather than just looking at its symptoms: however, to date, there is no clear consensus on how to fit the different levels together. In this review, we have described a possible method, based on a theoretical approach in which use is made of virtual laboratories concretely capable of implementing these notions: it is our belief that this approach could succeed in making sense of the deep mechanics that govern the underlying processes of the brain, thus helping neuroscientists in their daily work.

With such a complete model of the human brain, four main objectives could be addressed: i) earlier and more accurate detection of AD; ii) new surrogate outcomes for clinical trials; iii) faster development of drugs aimed at delaying or halting the neurodegeneration; iv) the development of a reference model that could be used also in other multimodal neurodegenerative brain diseases and research communities.

Finally, the idea of defining a multiscale and multimodal approach to further understanding of the complex pathophysiology of AD has recently been turned into major projects. In Europe, a billion-euro initiative, the Human Brain Project (HBP), is currently under way [12, 48, 8], wherein all existing knowledge about the human brain is to be pulled together in order to build up a model of the brain, piece by piece, with advanced ITC simulations. This idea has also prompted a similar initiative called the Brain Activity Map project (BAM), this time in the US. The ultimate aim of this project is to map the activity of every single neuron in the human brain. This historical moment will certainly be destined to leave a large footprint in our community and in the way we conduct (neuro)science. As Henry Makram, HBP principal investigator, has claimed: "It is not impossible to build a human brain and we can do it in 10 years" reconstructing deeply the brain's every circuit and process. Therefore, there is nothing left to do but work!

REFERENCES

1. Alzheimer A, Stelzmann RA, Schnitzlein HN, et al (1995). An English translation of Alzheimer's 1907 paper, "Uber eine eigenartige Erkankung der Hirnrinde". Clin Anat 8: 429-431.
2. Antila K, Lötjönen J, Thurfjell L, et al (2013). The PredictAD project: development of novel biomarkers and analysis software for early diagnosis of the Alzheimer's disease. Interface Focus 3: 20120072
3. Astarita G, Piomelli D (2011). Towards a whole-body systems [multi-organ] lipidomics in Alzheimer's disease. Prostaglandins Leukot Essent Fatty Acids 85:197-203.
4. Barreto P de S (2013). Alzheimer's disease: learning from the past, looking to the future. Am J Alzheimers Dis Other Demen 28: 304-305.
5. Bencherif B, Endres CJ, Musachio JL, et al (2002). PET imaging of brain acetylcholinesterase using [11C]CP-126,998, a brain selective enzyme inhibitor. Synapse 45: 1-9.
6. Bernal-Rusiel JL, Greve DN, Reuter M, et al (2012). Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. Neuroimage 66C: 249-260.
7. Blennow K, Zetterberg H (2013). The application of cerebrospinal fluid biomarkers in early diagnosis of Alzheimer disease. Med Clin North Am 97: 369-376.
8. Calimera A, Macii E, Poncino M (2013) The Human Brain Project and neuromorphic computing. Funct Neurol 28:
9. Chaari L, Pesquet J.-C., Tourneret J.-Y., et al (2010). A hierarchical Bayesian model for frame representation. IEEE Transactions on Signal Processing 58: 5560-5571.
10. Clark CM, Pontecorvo MJ, Beach TG, et al (2012). Cerebral PET with florbetapir compared with neuropathology at autopsy for detection of neuritic amyloid-beta plaques: a prospective cohort study. Lancet Neurol 11: 669-678.
11. Cordonnier C, van der Flier WM (2011). Brain microbleeds and Alzheimer's disease: innocent observation or key player? Brain 134: 335-344.
12. D'Angelo E, Solinas S, Garrido J, et al (2013). Realistic modeling of neurons and networks: towards brain simulation. Funct Neurol 28:
13. Dinov ID, Van Horn JD, Lozev KM, et al (2009). Efficient, distributed and interactive neuroimaging data analysis using the LONI pipeline. Front Neuroinform 3:22.
14. Egerházi A, Berecz R, Bartók E, et al (2007). Automated Neuropsychological Test Battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. Prog Neuropsychopharmacol Biol Psychiatry 31:746-751.
15. Ewers M, Frisoni GB, Teipel SJ, et al (2011). Staging Alzheimer's disease progression with multimodality neuroimaging. Prog Neurobiol 95: 535-546.
16. Fischl B (2012). FreeSurfer. NeuroImage 62: 774-781.
17. Flicek P, Ahmed I, Amode MR, et al (2013). Ensembl 2013. Nucleic Acids Res 41 (Database issue): D48-55.
18. Fox NC, Warrington EK, Freeborough PA, et al (1996). Presymptomatic hippocampal atrophy in Alzheimer's disease. A longitudinal MRI study. Brain 119: 2001-2007.
19. Frisoni GB, Fox NC, Jack CR Jr, et al (2010). The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6:67-77.
20. Frisoni GB, Redolfi A, Manset D, et al (2011). Virtual imaging laboratories for marker discovery in neurodegenerative diseases. Nat Rev Neurol 7: 429-438.
21. Frisoni GB, Winblad B, O'Brien JT (2011). Revised NIA-AA criteria for the diagnosis of Alzheimer's disease: a step forward but not yet ready for widespread clinical use. Int Psychogeriatr 23: 1191-1196.
22. Friston KJ (1995). Commentary and opinion: II. Statistical parametric mapping: ontology and current issues. J Cereb Blood Flow Metab 15: 361-370.

23. Friston KJ, Glaser DE, Henson RN, et al (2002a). Classical and Bayesian inference in neuroimaging: applications. Neuroimage 16: 484-512.
24. Friston KJ, Penny W, Phillips C, et al (2002b). Classical and Bayesian inference in neuroimaging: theory. Neuroimage 16: 465-483.
25. Furukawa K, Okamura N, Tashiro M, et al (2010). Amyloid PET in mild cognitive impairment and Alzheimer's disease with BF-227: comparison to FDG-PET. J Neurol 257: 721-727.
26. Genovese CR, Lazar NA, Nichols T (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15: 870-878.
27. George-Hyslop PS, Rossor M (2001). Alzheimer's disease. Unravelling the disease process. Lancet 358 Suppl: S1.
28. Graff-Radford J, Kantarci K (2013). Magnetic resonance spectroscopy in Alzheimer's disease. Neuropsychiatr Dis Treat 9: 687-696.
29. Hampel H, Teipel SJ, Fuchsberger T, et al (2004). Value of CSF beta-amyloid1-42 and tau as predictors of Alzheimer's disease in patients with mild cognitive impairment. Mol Psychiatry 9:705-710.
30. Herholz K, Westwood S, Haense C, et al (2011). Evaluation of a calibrated (18)F-FDG PET score as a biomarker for progression in Alzheimer disease and mild cognitive impairment. J Nucl Med 52: 1218-1226.
31. Herskovits EH, Peng H, Davatzikos C (2004). A Bayesian morphometry algorithm. IEEE Trans Med Imaging 23: 723-737.
32. Hochberg Y, Benjamini Y (1990). More powerful procedures for multiple significance testing. Stat Med 9: 811-818.
33. Jack CR Jr (2012). Alzheimer disease: new concepts on its neurobiology and the clinical role imaging will play. Radiology 263: 344-361.
34. Jack CR Jr, Knopman DS, Jagust WJ, et al (2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol 12: 207-216.
35. Jack CR Jr, Knopman DS, Jagust WJ, et al (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 9: 119-128.
36. Jack CR Jr, Knopman DS, Weigand SD, et al (2012). An operational approach to National Institute on Aging-Alzheimer's Association criteria for preclinical Alzheimer disease. Ann Neurol 71:765-775.
37. Jack CR Jr, Lowe VJ, Senjem ML, et al (2008). 11C PiB and structural MRI provide complementary information in imaging of Alzheimer's disease and amnestic mild cognitive impairment. Brain 131: 665-680.
38. Johansson P, Mattsson N, Hansson O, et al (2011). Cerebrospinal fluid biomarkers for Alzheimer's disease: diagnostic performance in a homogeneous mono-center population. J Alzheimers Dis 24:537-546.
39. Karran E, Mercken M, De Strooper B (2011).The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. Nat Rev.Drug Discov 10: 698-712.
40. Kim JS, Singh V, Lee JK, et al (2005). Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. Neuroimage 27: 210-221.
41. Klunk WE (2011) Amyloid imaging as a biomarker for cerebral beta-amyloidosis and risk prediction for Alzheimer dementia. Neurobiol Aging 32 (Suppl 1): S20-36.
42. Klunk WE, Mathis CA (2008). The future of amyloid-beta imaging: a tale of radionuclides and tracer proliferation. Curr Opin Neurol 21:683-687.
43. Leung KK, Barnes J, Modat M, et al (2011). Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. Neuroimage 55:1091-1108.
44. Leung KK, Ridgway GR, Ourselin S, et al (2012). Consistent multi-time-point brain atrophy estimation from the boundary shift integral. Neuroimage 59: 3995-4005.
45. López-Arrieta JM, Schneider L (2006) Metrifonate for Alzheimer's disease. Cochrane Database Syst Rev (2):CD003155.

46. Lötjönen J, Wolz R, Koikkalainen J, et al (2011). Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. Neuroimage 56: 185-196.
47. Mangialasche F, Solomon A, Winblad B, et al (2010). Alzheimer's disease: clinical trials and drug development. Lancet Neurol 9:702-716.
48. Markram H (2013) Seven challenges of neuroscience. Funct Neurol 28*:*
49. McKhann GM, Knopman DS, Chertkow H, et al (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7:263-269.
50. Mullikin JC, Hunt SE, Cole CG, et al (2000). An SNP map of human chromosome 22. Nature 407: 516-520.
51. Nathoo FS, Lesperance M, Lawson A, et al (2012). Comparing variational Bayes with Markov chain Monte Carlo for Bayesian computation in neuroimaging. Stat Methods Med Res 22: 398-423.
52. Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapp 15: 1-25.
53. Pepe MS, Feng Z, Janes H, et al (2008). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst 100:1432-1438.
54. Petersen RC (2012) New clinical criteria for the Alzheimer's disease spectrum. Minn Med 95: 42-45.
55. Petrella JR, Mattay VS, Doraiswamy PM (2008) Imaging genetics of brain longevity and mental wellness: the next frontier? Radiology 246: 20-32.
56. Pillai JA, Cummings JL (2013) Clinical trials in predementia stages of Alzheimer disease. Med Clin North Am 97: 439-457.
57. Qizilbash N, Birks J, Lopez Arrieta J, et al (2007). WITHDRAWN: Tacrine for Alzheimer's disease. Cochrane Database Syst Rev (3):CD000202.
58. Ramanan VK, Risacher SL, Nho K, et al (2013). APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. Mol Psychiatry doi: 10.1038/mp.2013.19.
59. Ridha BH, Barnes J, Bartlett JW, et al (2006). Tracking atrophy progression in familial Alzheimer's disease: a serial MRI study. Lancet Neurol 5: 828-834.
60. Roberson ED, Mucke L (2006). 100 years and counting: prospects for defeating Alzheimer's disease. Science 314: 781-784.
61. Scheinin NM, Aalto S, Koikkalainen J, et al (2009). Follow-up of [11C]PIB uptake and brain volume in patients with Alzheimer disease and controls. Neurology 73: 1186-1192.
62. Shin J, Kepe V, Barrio JR, et al (2011). The merits of FDDNP-PET imaging in Alzheimer's disease. J Alzheimers Dis 26 (Suppl 3):135-145.
63. Silvestri L, Sacconi L, Pavone FS (2013). The connectomics challenge. Funct Neurol 28:
64. Smith AD (2002). Imaging the progression of Alzheimer pathology through the brain. Proc Natl Acad Sci U S A 99: 4135-4137.
65. Soininen H, Mattila J, Koikkalainen J, et al (2012). Software tool for improved prediction of Alzheimer's disease. Neurodegener Dis 10: 149-152.
66. Sperling R, Johnson K (2013). Biomarkers of Alzheimer disease: current and future applications to diagnostic criteria. Continuum (Minneap Minn) 19: 325-338.
67. Teipel SJ, Sabri O, Grothe M, et al (2013). Perspectives for multimodal neurochemical and imaging biomarkers in Alzheimer's disease. J Alzheimers Dis 33 (Suppl 1): S329-347.
68. Vellas B, Aisen PS (2010). Early Alzheimer's trials: new developments. J Nutr Health Aging 14: 293.
69. Vercauteren T, Pennec X, Perchant A, et al (2009). Diffeomorphic demons: efficient non-parametric image registration. Neuroimage 45, (1 Suppl): S61-72.

70. Wakefield J, De Vocht F, Hung RJ (2010) Bayesian mixture modeling of gene-environment and gene-gene interactions. Genetic Epidemiol 34:16-25.
71. Wolz R, Aljabar P, Hajnal JV, et al (2010). LEAP: learning embeddings for atlas propagation. Neuroimage 49: 1316-1325.
72. Woolrich MW, Jbabdi S, Patenaude B, et al (2009). Bayesian analysis of neuroimaging data in FSL. Neuroimage 45 (1Suppl): S173-186.
73. Zhang H, Awate SP, Das SR, et al (2010). A tract-specific framework for white matter morphometry combining macroscopic and microscopic tract features.Med Image Anal 14: 666-673.

Chapter 3

Head-to-head comparison of two popular cortical thickness extraction algorithms: a cross-sectional and longitudinal study

Alberto Redolfi
David Manset
Frederik Barkhof
Lars-Olof Wahlund
Tristan Glatard
Jean-François Mangin
Giovanni B Frisoni
NeuGRID Consortium
Alzheimer's Disease Neuroimaging Initiative

PLoS One
2015 Mar 17;10(3):e0117692.

ABSTRACT

Background and purpose

The measurement of cortical shrinkage is a candidate marker of disease progression in Alzheimer's. This study evaluated the performance of two pipelines: Civet-CLASP (v1.1.9) and Freesurfer (v5.3.0).

Methods

Images from 185 ADNI1 cases (69 elderly controls (CTR), 37 stable MCI (sMCI), 27 progressive MCI (pMCI), and 52 Alzheimer (AD) patients) scanned at baseline, month 12, and month 24 were processed using the two pipelines and two interconnected e-infrastructures: neuGRID (https://neugrid4you.eu) and VIP (http://vip.creatis.insa-lyon.fr). The vertex-by-vertex cross-algorithm comparison was made possible applying the 3D gradient vector flow (GVF) and closest point search (CPS) techniques.

Results

The cortical thickness measured with Freesurfer was systematically lower by one third if compared to Civet's. Cross-sectionally, Freesurfer's effect size was significantly different in the posterior division of the temporal fusiform cortex. Both pipelines were weakly or mildly correlated with the Mini Mental State Examination score (MMSE) and the hippocampal volumetry. Civet differed significantly from Freesurfer in large frontal, parietal, temporal and occipital regions ($p<0.05$). In a discriminant analysis with cortical ROIs having effect size larger than 0.8, both pipelines gave no significant differences in area under the curve (AUC). Longitudinally, effect sizes were not significantly different in any of the 28 ROIs tested. Both pipelines weakly correlated with MMSE decay, showing no significant differences. Freesurfer mildly correlated with hippocampal thinning rate and differed in the supramarginal gyrus, temporal gyrus, and in the lateral occipital cortex compared to Civet ($p<0.05$). In a discriminant analysis with ROIs having effect size larger than 0.6, both pipelines yielded no significant differences in the AUC.

Conclusions

Civet appears slightly more sensitive to the typical AD atrophic pattern at the MCI stage, but both pipelines can accurately characterize the topography of cortical thinning at the dementia stage.

INTRODUCTION

Structural imaging has had a long role as biomarker of progression among entry criteria for AD trials [1]. The advent of disease-modifying therapies has led to interest in the use of magnetic resonance imaging (MRI) as a possible "surrogate" measure of outcome. The two most established markers of progression on MRI are the hippocampal and the whole brain atrophy rates [2]. However, the first study assessing the effects of β-amyloid immunotherapy reported surprising findings, i.e. greater hippocampal and whole-brain atrophy rates in patients treated with AN1792 vaccination [3]. On the contrary, cortical thickness might be a promising "global" measure of disease progression, as it could represent a marker more specifically related to the evolution of AD evolution [4,5] and might be useful to evaluate the efficacy of new disease-modifying therapies [6].

Several tools for the automatic extraction of cortical thickness have been developed, each based on different levels of complexity, robustness, and automation. Among others, the Civet-CLASP pipeline [7] and Freesurfer [8] are the two most exploited algorithms within the neuroscientific community. Obtaining an accurate thickness measurement requires the explicit reconstruction of the outer boundary on the base of the inner boundary [9], which can be done along two different approaches: (I) a skeleton method or (II) a model-based deformation of the inner surface. CIVET makes use of the skeleton mesh-based approach called constrained Laplacian anatomic segmentation using proximity. The pial surface is expanded from the white surface up to the boundary between gray matter and CSF, along a Laplacian map [10]. Terms for stretch and self-proximity are included to regularize the deforming mesh and avoid mesh self-intersection inside sulci. Differently, Freesurfer makes use of iterative and adaptive deformation and segmentation methods, deforming the mesh to reconstruct the inner and the pial surfaces. Freesurfer uses a routine function to find and correct the topological defects in the initial inner surface. The deformable model is constrained by a second-order smoothing term [11] and by a mesh self-intersection prevention routine [8], which both help to correctly establish the boundaries between adjacent banks in tight sulci. Unfortunately, some relevant problems hamper the use of these techniques. Both tools measure the cortical thickness from two 3D cortical sheets, each of which is composed by thousands of vertices and faces, making the reconstruction of the cortical mantle a complex and time consuming procedure [12].

Although several methods have been proposed in the past decades, little work has been done to compare their performances on real clinical datasets [13]. The aim of this study was to perform a head-to-head comparison between Civet-CLASP and Freesurfer. This can be

considered a mandatory step toward the standardization of cortical thickness biomarkers, which in turn will pave the way to effectively translate a three-dimensional cortical marker to innovative disease modifying trials.

MATERIALS AND METHODS

Subjects

The sample group we selected consisted of 185 subjects (69 normal elderly controls (CTR), 37 stable MCI (sMCI), 27 progressive MCI (pMCI), and 52 Alzheimer (AD) patients), belonging to the Alzheimer's Disease Neuroimaging Initiative (ADNI1). Demographics and clinical data are summarized in Table 1.

	CTR	sMCI	pMCI	AD	P
Number	69	37	27	52	
Age (y)	75.6 ± 4.8	74.6 ± 7.5	73.1 ± 8	76.0 ± 6.2	N.S.
Education (y)	15.9 ± 2.9	15.6 ± 3.3	16.6 ± 2.1	15.0 ± 2.6	N.S.
Gender (M/F)	38/31	22/15	17/10	27/25	0.002
MMSE (BSL)	29.2 ± 1.0	27.4 ± 2.0	27.1 ± 1.7	23.4 ± 2.3	<0.001
Δ MMSE	0.1 ± 1.4	- 0.4 ± 1.8	- 3.4± 3.6	- 3.9 ± 5.1	<0.001
CDR	0 (69)	0.5 (37)	0.5 (27)	0.5 (27)–1 (25)	<0.001
ApoE ε4 carriers (%)					
-/-	66.5	35	40.5	31	
-/+	28	60	44.5	50	
+/+	5.5	5	15	19	

Table1 - Demographic and clinical characteristics. Data are expressed as mean value ± standard deviation (σ). BSL: Baseline; Δ: Difference between month 24 and baseline; MMSE: Mini Mental State Examination scores; CDR: Clinical Dementia Ratings score; CTR: Controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's Disease; P: significance on Fisher's exact test or ANOVA; N.S.: not significant.

MMSE and CDR scores differed significantly among the four groups (P<0.001), while age and educational levels were not significantly different. There was a significant difference in sex (P < 0.002) with a higher prevalence of male. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI1 study is conducted in accordance with the Good Clinical Practice guidelines, the Declaration of Helsinki, and U.S. 21 CFR Part 50 (Protection of Human Subjects), and Part 56 (Institutional Review Boards). ADNI1 study was approved by the Institutional Review Boards (IRB) of all

of the participating institutions. Specifically, they are: Albany Medical College, Banner Alzheimer's Institute, Baylor College of Medicine, Boston University, Brigham and Women's Hospital, Butler Hospital Memory & Aging Program, Case Western Reserve University, Cleveland Clinic, Columbia University, Darthmouth – Hitchcock Medical Center, Dent Neurologic Institute, Duke University Medical Center, Emory University, Georgetown University, Howard University, Indiana University, Jefferson Hospital for Neuroscience, Johns Hopkins University, Mayo Clinic, Jacksonville, Mayo Clinic, Rochester, McGill University/Jewish General Hospital Memory Clinic, Medical University of South Carolina, Mount Sinai School of Medicine, Neurological Care of Central New York, New York University Medical Center, Northwestern University, Ohio State University, Olin Neuropsychiatry Research Center, Oregon Health and Science University, Parkwood Hospital, Premiere Research Institute, Rhode Island Hospital, Rush University Medical Center, Saint Joseph's Health Center, Stanford University, Banner Sun Health Research Institute, Sunnybrook Health Sciences, University of Alabama, Birmingham, University of British Columbia, University of California, Davis, University of California, Irvine, University of California, Irvine-BIC, University of California - Los Angeles, University of California - San Diego, University of California - San Francisco, University of Kansas, University of Kentucky, University of Michigan, Ann Arbor, University of Nevada School of Medicine, Las Vegas, University of Pennsylvania, University of Pittsburgh, University of Rochester, University of Southern California, University of Texas Southwestern Medical Center, University of Wisconsin, Wake Forest University, Washington University St. Louis, Wein Center for Clinical Research and Yale University School of Medicine. Informed written consent was obtained from all participants at each site. A detailed description of the study procedures, IRB approval and informed written consents is available at http://www.adni-info.org/pdfs/adni_protocol_9_19_08.pdf (section D.5). Data used in this analysis were downloaded from the ADNI database (http://adni.loni.usc.edu/). List of subjects' RIDs can be found in supplementary S1 Table.

Research infrastructures and pipelines

The evaluation of the cortical thickness is a computationally demanding task. We used two online e-infrastructures, namely neuGRID (https://neugrid4you.eu) [14] and VIP (http://vip.creatis.insa-lyon.fr) [15] to massively distribute job analyses, thus reducing the overall processing. Civet's and Freesurfer's main features are summarized as follow:

MRI input file format	MINC	DICOM; NIFTI
Mesh output format	MNI OBJ	PIAL
3D mesh generation	Iterative morphing method (i.e.: skeleton-based reconstruction). The pial surface is expanded from the white surface to the boundary between gray matter and CSF along a Laplacian map	Iterative adaptive morphing and segmentation methods (i.e.: model-based deformation of the inner surface)
3D mesh pitfalls	Requiring corrections for topological errors	Affected by geometric inaccuracies
Sensitivity to artifacts	Variation of the signal intensity across the image; shading artifacts; intensity non-uniformity (through N3 procedure); poor radio frequency field uniformity, and eddy currents are mitigated	Motion correction (when there are multiple MR source volumes of the same subject) and non-uniform intensity normalization in MR data (through N3 procedure) are carried out
Cortex representation	Geometrically accurate	Topologically correct
Longitudinal stream	Option not available	Option available. Freesurfer, during the longitudinal stream, through repeated cross-sectional measures from the same subject reduces the variability of the cortical thickness estimation. In the present study the longitudinal stream has been used (i.e.: baseline, month 12, and month 24)
Computational time	≅7 hours per single subject	≅35 hours per single subject
Intra-algorithm thickness repeatability	High	High
	(no differences between MPRAGE and MPRAGE repeat FDR corrected p-maps)	(no differences between MPRAGE and MPRAGE repeat FDR corrected p-maps)
Number of vertices on the cortical surface	ICO6	ICO7
	(# Vertices = 2*(10*4^n+2) = 81′924 vertices)	(# Vertices = 2*(10*4^n+2) = 327′680 vertices)
Max image voxel resolution as input	0.5 x 0.5 x 0.5 mm^3	1.0 x 1.0 x 1.0 mm^3
Average thickness	1 mm thicker than Freesurfer in all the diagnostic groups and in all time points (baseline and month 24)	1 mm thinner than Civet in all the diagnostic groups and in all time points (baseline and month 24)
Strengths	*CROSS-SECTIONAL*	*CROSS SECTIONAL*
	1) Constant thinning progression in different disease stages	1) Thinning progression peaks earlier than Civet
	2) Weak to medium trend of correlation to both MMSE score and hippocampal volume	2) Slightly higher disease effect (Hedge's g) in comparing CTR with pMCI and AD
	3) Sensitive in expected cortical regions affected by disease neuropathology (i.e.: cingulate, dorsolateral frontal and parietal cortex)	3) Higher, but not significant, AUC to discriminate CTR versus pMCI or AD
		4) Sensitive in expected but also scattered unexpected cortical regions affected by disease neuropathology
	LONGITUDINAL	*LONGITUDINAL*
	1) Higher disease effect in pMCI and AD	1) Higher disease effect trend in CTR
	2) More sensitive to significant atrophic patterns in frontal-parietal regions (especially in pMCI)	2) Better correlation with hippocampal volumetric atrophy
	3) Sensitive to detect statistical significant atrophic differences between: AD vs CTR; AD vs sMCI; pMCI vs CTR	3) Sensitive to detect statistical significant atrophic differences between: AD vs CTR; AD vs sMCI
	4) Sensitive enough to detect statistical significant atrophic differences in many temporal ROIs between: sMCI vs pMCI	4) Higher, but not significant, AUC to discriminate pMCI due to AD in a time span of 2 years
Algorithm Validations	Manual method of tagging GM/CSF and GM/WM interfaces of forty brains on twenty regions of interest of young healthy volunteers (Kabani et al. 2001)	Comparison against post-mortem subjects with Huntington Disease and healthy control (Rosas et al. 2002) both of 43 years old

Table 2 – Comparative table. Comparative table where the main characteristics of the pipelines involved in this head-to-head comparison are summarized. MINC: Medical Imaging Network Common Data Form; DICOM: Digital Imaging and Communications in Medicine; NIFTI: Neuroimaging Informatics Technology Initiative; MNI OBJ: geometry file format developed by the Montreal Neurological Institute; PIAL: geometry file format developed by Martinos Center for Biomedical Imaging.

- Civet-CLASP uses an iterative morphing method and intensity non-uniformity correction; spatial normalization to stereotaxic space; tissue classification; cortical surface extraction; cortical thickness measurement. The correspondence among subjects is granted by the nonlinear registration of the sulcal geodesic depth map with an average sulcal depth sphere surface [10].
- Freesurfer uses iterative adaptative morphing/segmentation methods and relies on similar preprocessing steps, although differently arranged. The white matter derives from the segmentation and topology correction. Gray matter is derived along T1 intensity gradient. Correspondence among subjects is obtained through surface registration to the Freesurfer reference atlas. In this study, we used the longitudinal processing stream, where the variability is reduced using repeated measures from the same subject (i.e.: baseline, month 12 (data not shown), and month 24 cross-sectional analyses) as common information to initialize the process [16].

Table 2 reports the main features of the two pipelines.

Study design

The workflow of the study is reported as supplementary figure (see S1 Figure).

MRI acquisition

The Alzheimer's Disease Neuroimaging Initiative (ADNI) has a specific protocol for the acquisition and harmonization of MR images. The ADNI 3D T1-weighted structural images are acquired using selected systems from GE Healthcare, Philips Medical Systems and Siemens Medical Solutions, with an eye toward minimizing cross-platform differences. The Magnetization Prepared RApid Gradient Echo (MPRAGE) acquisition sequence has nominal T1 = 1000 ms, TR = 2400 ms and TE = 5 ms. The B2B acquisition set in ADNI1 is composed of a MPRAGE scan and a MPRAGE-repeat scan.

Visual quality control

All the post-processed scans output by neuGRID and VIP were quality controlled by an expert evaluator, who visually inspected them using the Matlab Imaging toolbox for 3D surfaces, which enables the user to rotate, zoom in and out the cortical surface along all the possible orientations. A reconstructed mesh was judged accurate when all the following 23 Sulci were visible and correctly reconstructed: (I) Sylvian Fissure, (II) Central Sulcus, (III) Postcentral

Sulcus, (IV) Precentral Sulcus, (V) Superior Temporal Sulcus, (VII) Intraparietal Sulcus, (VIII) Primary Intermediate Sulcus, (IX) Secondary Intermediate Sulcus, (X) Transverse Occipital Sulcus, (XI) Inferior Temporal Sulcus, (XII) Inferior Frontal Sulcus, (XII) Middle Frontal Sulcus, (XIV) Olfactory Sulcus, (XV) Occipital-Temporal Sulcus, (XVI) Collateral Sulcus, (XVII) Olfactory Control Line, (XVIII) Olfactory-Middle Frontal Control Line, (XIX) Middle Frontal-Precentral Control Line, (XX) Precentral-Central Control Line, (XXI) Central-Postcentral Control Line, (XXII) Postcentral-Transverse Occipital Control Line and (XXIII) Occipital Control Line. As a result of this visual QC, only one of the two B2B cortical surfaces was chosen for analyses.

Hybrid Template Generation enabling head-to-head (H2H) comparison

Cortex surfaces as extracted by Civet and Freesurfer are morphologically and topographically different. For an accurate comparison to be possible, it was necessary to deform the surface morphology of at least one algorithm. To map each point of one surface onto the other, we adopted an elastic non-rigid registration to get the right displacement vector. To our knowledge, Gradient Vector Flow (GVF) has not been used before to control 3D free form deformation. The vector field computed via GVF provided the directions along which each vertex of our source surface could evolve to match a corresponding point on the target surface. Once registered, space coordinates of each face vertices are coincident and vertices are spatially aligned. Subsequently, in order to compare the correct cortical index value at each vertex, we adopted the Closest Point Search (CPS) technique, essential to establish the correct topographical match of the same morphological points obtained with 3D GVF. For each point, CPS returned the mutual match between Civet's and Freesurfer's cortical thickness array. The entire process enabling the head-to-head comparison is illustrated in Figure 1. The procedure was implemented using Matlab (v2009b). The data generated in this study are made publicly available to promote the evaluation of cortical thickness tool (https://neugrid4you.eu/datasets).

Atlases and ROIs Definition

The head-to-head comparison and the ROI analyses between pipelines were done using the Harvard-Oxford cortical structural atlas. We chose 28 out of the 48 cortical areas provided [17], consistently with those used by other reference work groups [18-21]. For a complete list of the selected ROIs, see Table 3.

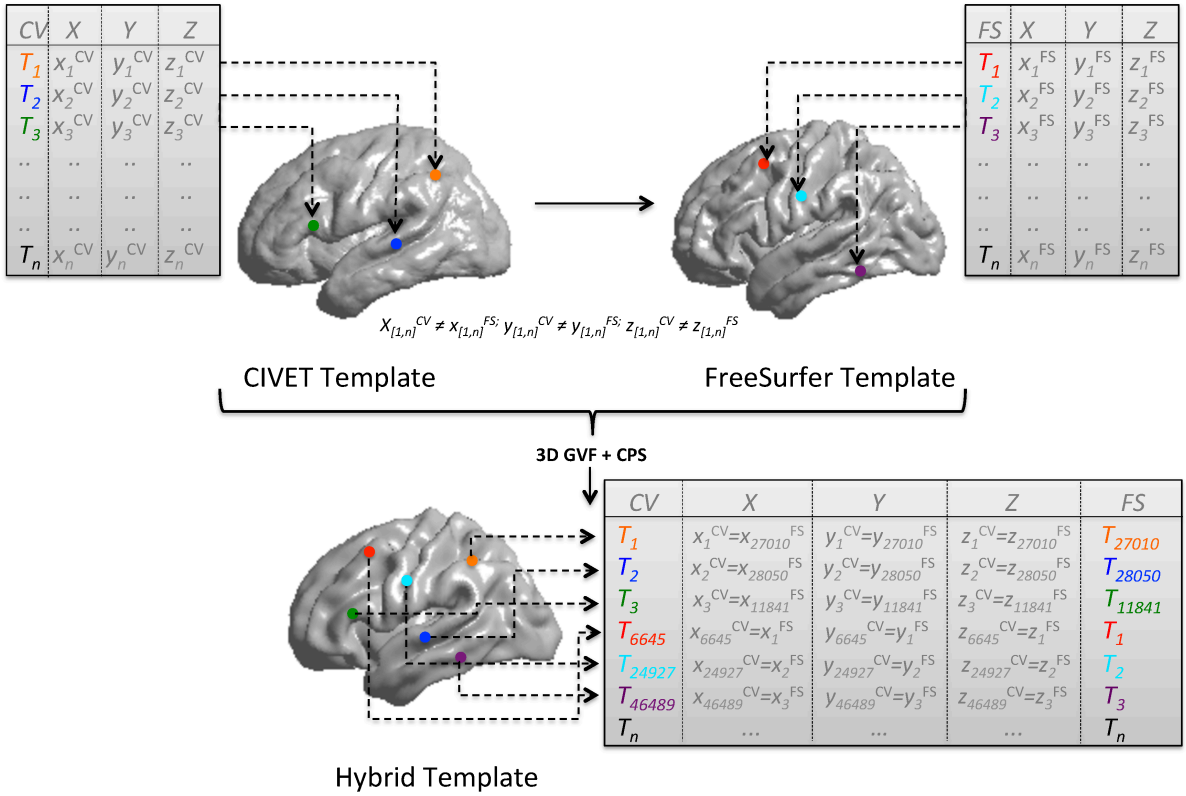


Figure 1 - Registration of templates and surface points correspondence. Source template is Civet's surface while target template is the Freesurfer' surface template. Starting from two averaged surfaces (previously created from the same set of 10 CTR, 10 sMCI, and 10 AD brains) the hybrid template (characterized by 81924 vertices and 163840 faces) is derived after 15 GFV iterations. In GVF, deformations are achieved by tuning an underlying set of control points (187×187×187) in the source surface. Control point displacements are then interpolated to obtain a continuous transformation through basis spline functions. To keep the contour smooth, a membrane and percentage thin plate energy was used as regularization. The parameters defining the attraction to edges and energy surfaces were empirically determined. Finally, the CPS step defined the mutual correspondence of Civet and Freesurfer thickness values for each vertex. CV: Civet; FS: Freesurfer; X-Y-Z: value of the vertex space coordinates; T: value of the cortical thickness for each vertex; n: number of vertices (min=0; max=81924); 3D GVF: 3D gradient vector flow; CPS: Closest point search.

Statistical analysis to compare Cortical Thinning patterns

Cortical thinning within the same diagnostic groups was assessed using paired samples t-tests. P-maps were corrected for multiple comparisons using the False Discovery Rate (FDR; α=0.01) method [22]. Tukey-Kramer post-hoc testing of ANOVA ($\alpha = 0.05$ in cross sectional comparison and $\alpha = 0.01$ in longitudinal analysis) was used to test thinning differences among

the diagnostic groups and the different ROIs analyzed. Effect sizes were computed as Hedge's g and Z-tests were performed to assess significant discrepancies between the performances of each pipeline. Correlations of cortical thickness to MMSE scores and hippocampal volumes were investigated, Steiger's Z was used to assess significant differences between Pearson's r values. Logistic regressions were applied on pre-selected thickness ROIs, and Receiver Operating Characteristic (ROC) curves were used to assess discriminative accuracy of the two pipelines. AUCs were statistically compared using the method adopted by Hanley and McNeil [23], setting the threshold for significance at a p value of 0.05. Kendall's tau coefficients were calculated and the derived z-test converted into the Pearson's correlation coefficient. Statistical analysis was performed with Matlab (v2009b).

Cortical Metrics

Both pipelines define thickness as the Euclidean distance and both can produce maps not restricted to the original MRI voxel resolution: thus, they can detect sub-millimeter differences between and within groups [8,24]. For the sake of this article, we defined the concept of "disease effect" as the relative predominance of one pipeline over the other to detect atrophy when comparing two groups (G) or two time-points (T):

$$\boldsymbol{DISEASE\ EFFECT}\ \{\Delta \boldsymbol{G} \mid \Delta \boldsymbol{T}\} = \left(\Delta^{\boldsymbol{FS}}\ \boldsymbol{Thickness} - \Delta^{\boldsymbol{CV}}\ \boldsymbol{Thickness}\right) \quad \textbf{(1)}$$

The values of the disease effect are mapped vertex by vertex on the hybrid template previously created (see Figure 2 and 3 panel b).

RESULTS

Comparison of cortical metrics

The reconstruction of cortical thickness from B2B scans provided identical outcomes within the same pipeline (see supplementary S2 Figure).

Compared to Civet, Freesurfer provided absolute values systematically lower by about 30% (see supplementary S3 Figure). The difference between Civet and Freesurfer with respect to between-subjects variability (CoV) [25] ranges between 17-26% in the different diagnostic groups. The whole cortical thickness value at baseline and at month 24 is reported as supplementary S2 Table; both Civet and Freesurfer showed increasing values of thinning rates with the progression of the pathology. The relative percentage of thinning in paired diagnostic groups at baseline is reported as supplementary S3 Table; no statistical differences among the groups were detected in neither pipelines. The percentage of longitudinal thinning rate across the four different diagnostic groups is reported as S4 Table; both pipelines detected differences between AD versus CTR, and between AD versus sMCI; moreover, Civet was able to detect a significant longitudinal thinning difference between pMCI versus CTR.

Cross-sectional and longitudinal thinning differences between Civet and Freesurfer

Figure 2 compares CTR with sMCI, pMCI, and AD at baseline, and shows the details of the differences between Civet and Freesurfer at the individual vertex level. Figure 3 compares, for each diagnostic group, the longitudinal (2 years) cortical thinning rate at the individual vertex level as computed by Civet and Freesurfer.

ROI Analysis

Table 3 represents the comparison of the cross-sectional thickness differences at baseline, while Table 4 represents the longitudinal thinning rates with respect to the 28 selected ROIs. Cross-sectionally, the multiple comparison procedure highlighted small differences. Civet indicated as significant the temporal planum ROI, while Freesurfer identified as significant the superior parietal lobe. Longitudinally, Civet appeared to be much more sensitive in detecting significant thinning rate differences between CTR and AD in all the 28 ROIs considered, as opposed to only 22 ROIs as detected by Freesurfer (check symbol ¥). Comparing sMCI to AD, Civet was able to detect significant longitudinal thinning rate changes in all the 28 ROIs,). Again, Civet was able to detect significant longitudinal thinning rate changes between CTR and pMCI in 18 ROIs, as opposed to only 10 ROIs in Freesurfer (check symbol ¢). Lastly,

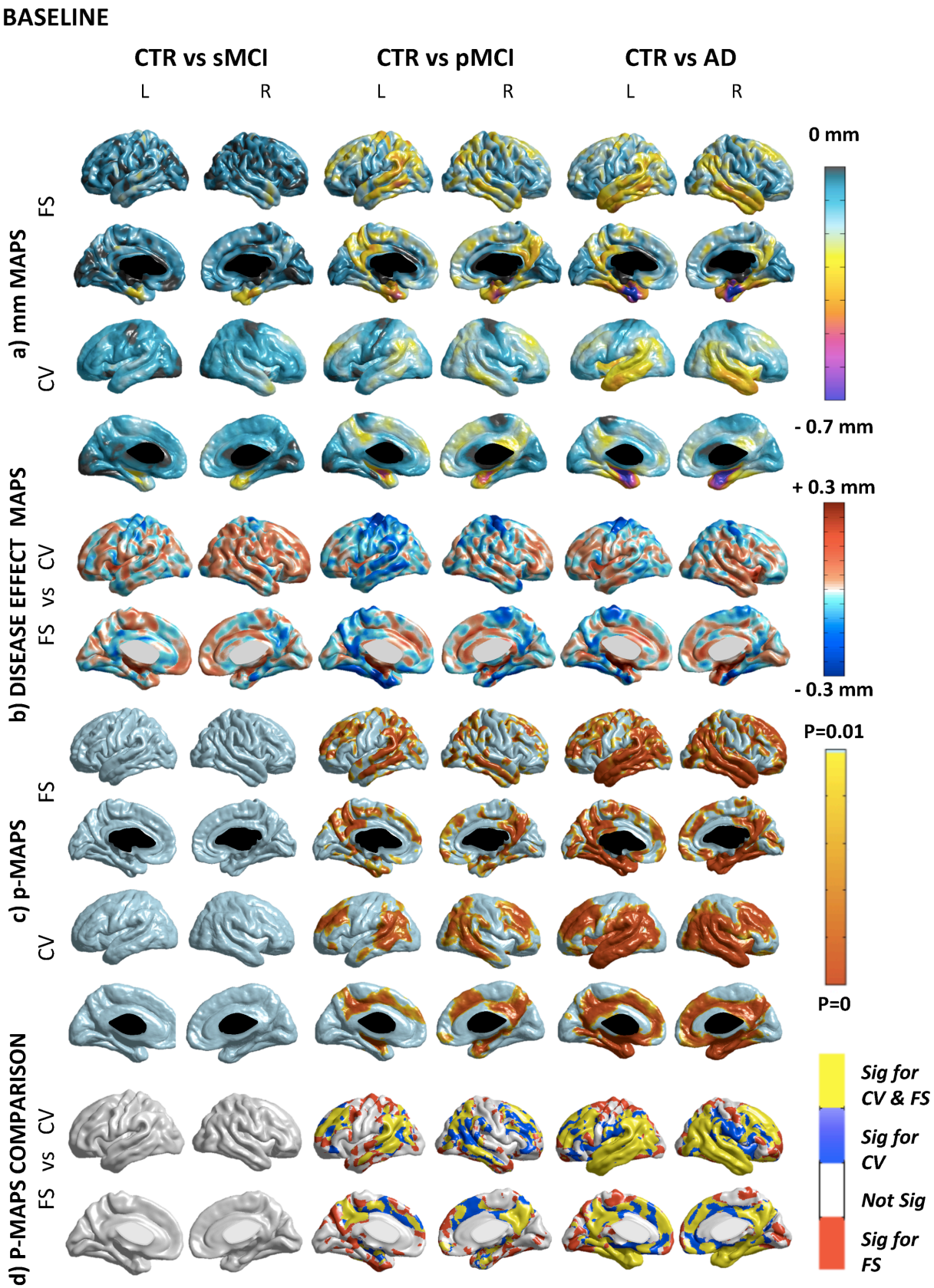


Figure 2 - Cross-sectional comparison.

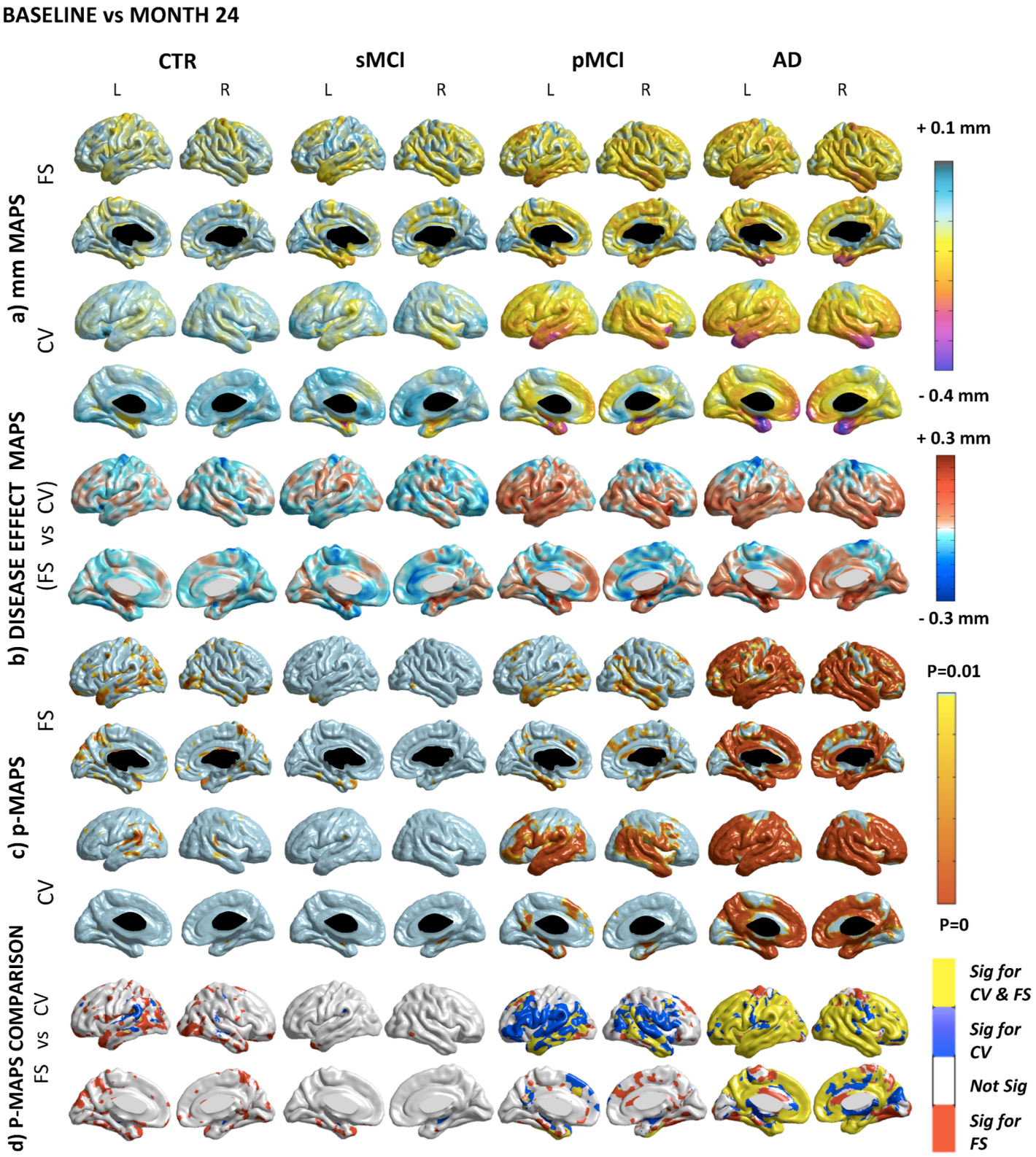


Figure 3 - Longitudinal comparison.

Figure 2 - Cross-sectional comparison: A) Absolute difference maps (mm) in Freesurfer and Civet. The degree of atrophy ranges between 0.1 and 0.7 mm in the different areas of the cortical mantle. B) Disease effect maps. There is a consistent delta (±0.3 mm) among the compared groups. Negative value means higher disease effect for Freesurfer (i.e.: parietal-temporal and precuneus areas); positive value means higher disease effect for Civet (i.e.: association areas and limbic parts of the cortex). C) Statistical difference maps ($p<0.01$ FDR-corrected). No significant voxels were found comparing CTR to sMCI. Atrophic areas were found contrasting pMCI with CTR (i.e.: the posterior cingulate, temporal lobe and frontal gyrus) with both tools. Comparing CTR versus AD the statistical significance extended (i.e.: medial temporal, retrosplenial, and lateral temporal regions). D) Overlapping and not-overlapping atrophic regions are shown. Significant voxels detected by both pipelines are in yellow; voxels detected only by Civet are in blue; voxels detected only by Freesurfer are in red. CV: Civet; FS: Freesurfer; L: Left hemisphere; R: Right hemisphere; CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's Disease.

Figure 3 - Longitudinal comparison: A) Absolute difference maps (mm) in each group. In CTR and sMCI, both pipelines report a very mild and widespread cortical thinning rate in the motor, somatosensory, verbal and visual association cortex. In pMCI, the atrophy peaks at rates around 0.3 mm in the medial temporal cortex, temporal-parietal-frontal neocortices, with sparing of the sensorimotor strip and of the visual cortex. In AD, the atrophy in the same areas accelerates beyond 0.4 mm. B) Disease effect maps. The mean estimate of the longitudinal disease effect in CTR and sMCI as computed by Freesurfer is greater, although Civet shows higher results in few scattered areas. Furthermore, in the entire disease spectrum, Freesurfer exhibited higher disease effect in the motor cortex. In pMCI, Civet exhibits a greater disease effect except for the cingulate gyrus, while in the AD group the exception is represented by the precuneus. C) Statistical difference maps ($p<0.01$ FDR-corrected). In CTR, Civet detects an atrophic cluster in the angular gyrus; while Freesurfer in the precuneus and in the temporo-occipital lobe. The pattern in sMCI was more reduced than in CTR. In pMCI Freesurfer was not able to find many regions detected by Civet with the same significance and extension (i.e.: orbital, triangulal, and opercular portion of the inferior frontal gyrus, transverse-temporal and mesial part of the superior frontal cortex, inferior parietal cortex, the superior temporal gyrus). Freesurfer was more sensitive in few scattered expected and unexpected regions. For both pipelines, the longitudinal AD shrinkage showed significant areas throughout the temporal, frontal and parietal lobes, consistently with the progression of the disease. Some shrivelling differences were detected in the anterior division of the cingulate, in the limbic lobe and in the cuneus. D) Overlapping and not-overlapping atrophic regions are shown. Significant voxels detected by both pipelines are in yellow; voxels detected only by Civet are in blue; voxels detected only by Freesurfer are in red. CV: Civet; FS: Freesurfer; L: Left hemisphere; R: Right hemisphere; CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's Disease.

BASELINE																
ROI			CIVET							FREESURFER						
			CTR vs sMCI		CTR vs pMCI		CTR vs AD		ANOVA	CTR vs sMCI		CTR vs pMCI		CTR vs AD		ANOVA
			Δ MEAN (mm) ± σ						P-value	Δ MEAN (mm) ± σ						P-value
3	Frontal	Superior Frontal Gyrus	-0.10	0.36	-0.18	0.38	-0.17	0.34	N.S.	-0.09	0.32	-0.21	0.35	-0.19	0.33	N.S.
4	Frontal	Middle Frontal Gyrus	-0.12	0.27	-0.18	0.29	-0.20	0.24	N.S.	-0.08	0.65	-0.20	0.68	-0.19	0.57	N.S.
5	Frontal	Inferior Frontal Gyrus, pars triangularis	-0.10	0.41	-0.13	0.35	-0.13	0.34	N.S.	-0.06	0.36	-0.13	0.37	-0.13	0.37	N.S.
6	Frontal	Inferior Frontal Gyrus, pars opercularis	-0.08	0.32	-0.12	0.26	-0.14	0.28	N.S.	-0.05	0.22	-0.15	0.23	-0.16	0.23	N.S.
33	Frontal	Frontal Orbital Cortex	-0.10	0.41	-0.13	0.39	-0.22	0.37	N.S.	-0.08	0.38	-0.13	0.35	-0.15	0.38	N.S.
18	Parietal	Superior Parietal Lobule	-0.09	0.47	-0.16	0.48	-0.14	0.45	N.S.	-0.06	0.23	-0.21 Π	0.20	-0.17 Ω	0.22	0.050
19	Parietal	Supramarginal Gyrus, anterior division	-0.07	0.30	-0.14	0.29	-0.19	0.28	N.S.	-0.05	0.32	-0.18	0.31	-0.16	0.30	N.S.
20	Parietal	Supramarginal Gyrus, posterior division	-0.10	0.27	-0.17	0.26	-0.23	0.27	N.S.	-0.08	0.33	-0.22	0.29	-0.21	0.33	N.S.
31	Parietal	Precuneus Cortex	-0.08	0.43	-0.18	0.37	-0.18	0.38	N.S.	-0.10	0.41	-0.25	0.40	-0.20	0.37	N.S.
22	Ocp.	Lateral Occipital Cortex, superior division	-0.09	0.36	-0.18	0.35	-0.18	0.41	N.S.	-0.06	0.35	-0.23	0.34	-0.21	0.40	N.S.
23	Ocp.	Lateral Occipital Cortex, inferior division	-0.07	0.29	-0.15	0.32	-0.19	0.25	N.S.	-0.05	0.30	-0.17	0.30	-0.20	0.27	N.S.
29	Lmb.	Cingulate Gyrus, anterior division	-0.05	0.47	-0.11	0.44	-0.11	0.45	N.S.	-0.06	0.48	-0.08	0.50	-0.08	0.53	N.S.
30	Lmb.	Cingulate Gyrus, posterior division	-0.08	0.54	-0.21	0.55	-0.22	0.51	N.S.	-0.10	0.65	-0.23	0.66	-0.21	0.62	N.S.
8	Temporal	Temporal Pole	-0.17	0.55	-0.18	0.55	-0.38	0.67	N.S.	-0.14	0.73	-0.24	0.69	-0.33	0.71	N.S.
9	Temporal	Superior Temporal Gyrus, anterior division	-0.10	0.50	-0.10	0.51	-0.23	0.59	N.S.	-0.07	0.99	-0.15	0.83	-0.20	0.90	N.S.
10	Temporal	Superior Temporal Gyrus, posterior division	-0.14	0.45	-0.15	0.38	-0.24	0.44	N.S.	-0.12	0.62	-0.19	0.63	-0.23	0.61	N.S.
11	Temporal	Middle Temporal Gyrus, anterior division	-0.13	0.51	-0.16	0.51	-0.31	0.48	N.S.	-0.12	0.74	-0.16	0.70	-0.25	0.70	N.S.
12	Temporal	Middle Temporal Gyrus, posterior division	-0.13	0.29	-0.18	0.23	-0.33	0.25	N.S.	-0.11	0.61	-0.20	0.49	-0.28	0.58	N.S.
13	Temporal	Middle Temporal Gyrus, temporo occipital part	-0.11	0.32	-0.20	0.35	-0.27	0.33	N.S.	-0.08	0.54	-0.23	0.54	-0.26	0.48	N.S.
14	Temporal	Inferior Temporal Gyrus, anterior division	-0.11	0.59	-0.14	0.62	-0.30	0.56	N.S.	-0.11	0.81	-0.17	0.87	-0.25	0.74	N.S.
15	Temporal	Inferior Temporal Gyrus, posterior division	-0.09	0.57	-0.15	0.59	-0.29	0.66	N.S.	-0.11	0.41	-0.21	0.41	-0.29	0.39	N.S.
16	Temporal	Inferior Temporal Gyrus, temporo occipital part	-0.08	0.64	-0.13	0.57	-0.21	0.58	N.S.	-0.09	0.62	-0.18	0.63	-0.24	0.63	N.S.
34	Temporal	Parahippocampal Gyrus, anterior division	-0.19	0.67	-0.28	0.67	-0.55	0.63	N.S.	-0.21	1.21	-0.33	1.22	-0.59	1.16	N.S.
35	Temporal	Parahippocampal Gyrus, posterior division	-0.11	0.57	-0.15	0.59	-0.32	0.56	N.S.	-0.10	1.25	-0.15	1.22	-0.25	1.23	N.S.
37	Temporal	Temporal Fusiform Cortex, anterior division	-0.14	0.57	-0.17	0.59	-0.35	0.52	N.S.	-0.15	0.80	-0.24	0.77	-0.41	0.73	N.S.
38	Temporal	Temporal Fusiform Cortex, posterior division	-0.11	0.37	-0.11	0.38	-0.30	0.35	N.S.	-0.13	0.46	-0.21	0.44	-0.34	0.44	N.S.
45	Temporal	Heschl's Gyrus (includes H1 and H2)	-0.09	0.18	-0.12	0.18	-0.22	0.18	N.S.	-0.11	0.24	-0.15	0.26	-0.17	0.25	N.S.
46	Temporal	Temporal Planum	-0.11	0.16	-0.13	0.14	-0.21 Ω	0.14	0.041	-0.11	0.37	-0.17	0.38	-0.20	0.37	N.S.

Table 3 - Cross sectional ROI-based analysis. Cross-sectional average cortical thinning differences (mm), standard deviation (σ), and Tukey-Kramer multiple comparison post-hoc analysis in ANOVA (P). The data refer to three groups: (a) CTR versus sMCI, (b) CTR versus pMCI and (c) CTR versus AD; α=0.05 level. Ω: Significant difference between "CTR versus sMCI" and "CTR versus AD"; Π: Significant difference between "CTR versus sMCI" and "CTR versus pMCI". N.S.: Not significant; CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's disease.

		ROI	BASELINE VS MONTH24																	
			CIVET									FREESURFER								
			CTR		sMCI		pMCI		AD		ANOVA	CTR		sMCI		pMCI		AD		ANOVA P-value
			Δ MEAN (mm) ± σ								P-value	Δ MEAN (mm) ± σ								
3	Frontal	Superior Frontal Gyrus	0.00	0.08	-0.04	0.12	-0.10	0.12	-0.13 ¥ ⊙	0.12	<0.0001	-0.04	0.10	-0.06	0.13	-0.12	0.14	-0.13 ¥	0.12	0.0002
4		Middle Frontal Gyrus	-0.03	0.06	-0.05	0.10	-0.11 ¢	0.12	-0.13 ¥ ⊙	0.12	<0.0001	-0.04	0.08	-0.06	0.13	-0.09	0.12	-0.11 ¥	0.13	0.0031
5		Inferior Frontal Gyrus, pars triangularis	-0.03	0.08	-0.03	0.11	-0.12 ¢ Ξ	0.11	-0.15 ¥ ⊙	0.13	<0.0001	-0.04	0.08	-0.05	0.15	-0.07	0.10	-0.09	0.11	N.S.
6		Inferior Frontal Gyrus, pars opercularis	-0.02	0.07	-0.04	0.10	-0.11 ¢	0.10	-0.14 ¥ ⊙	0.12	<0.0001	-0.05	0.11	-0.05	0.15	-0.07	0.12	-0.11	0.11	N.S.
#		Frontal Orbital Cortex	-0.12	0.10	-0.03	0.15	-0.11 ¢	0.14	-0.15 ¥ ⊙	0.16	<0.0001	-0.04	0.07	-0.05	0.11	-0.08	0.11	-0.10 ¥	0.09	0.0011
#	Parietal	Superior Parietal Lobule	-0.03	0.09	-0.02	0.10	-0.06	0.11	-0.09 ¥ ⊙	0.10	0.0006	-0.05	0.10	-0.05	0.14	-0.06	0.11	-0.08	0.10	N.S.
#		Supramarginal Gyrus, anterior division	-0.03	0.65	-0.04	0.09	-0.10 ¢	0.10	-0.12 ¥ ⊙	0.11	<0.0001	-0.05	0.11	-0.04	0.10	-0.09	0.13	-0.12 ¥	0.12	0.0022
#		Supramarginal Gyrus, posterior division	-0.04	0.07	-0.05	0.10	-0.12 ¢	0.10	-0.14 ¥ ⊙	0.11	<0.0001	-0.05	0.09	-0.05	0.12	-0.09	0.12	-0.12 ¥ ⊙	0.11	0.0010
#		Precuneus Cortex	-0.01	0.07	-0.02	0.07	-0.07	0.11	-0.11 ¥ ⊙	0.11	<0.0001	-0.05	0.08	-0.03	0.07	-0.09	0.10	-0.11 ¥ ⊙	0.10	<0.0001
#	Ocp.	Lateral Occipital Cortex, superior division	-0.03	0.07	-0.02	0.09	-0.09 Ξ	0.09	-0.11 ¥ ⊙	0.11	<0.0001	-0.04	0.10	-0.06	0.13	-0.12	0.14	-0.13 ¥	0.12	0.0002
#		Lateral Occipital Cortex, inferior division	-0.03	0.09	-0.03	0.11	-0.12 ¢	0.11	-0.15 ¥ ⊙	0.15	<0.0001	-0.06	0.07	-0.05	0.10	-0.10	0.10	-0.07	0.09	N.S.
#	Lmb.	Cingulate Gyrus, anterior division	0.00	0.08	0.01	0.13	-0.04	0.12	-0.11 ¥ ⊙	0.14	<0.0001	-0.02	0.07	-0.04	0.11	-0.07	0.08	-0.08 ¥	0.09	0.0024
#		Cingulate Gyrus, posterior division	-0.01	0.10	-0.02	0.07	-0.07	0.16	-0.12 ¥ ⊙	0.12	<0.0001	-0.04	0.06	-0.04	0.07	-0.10 ¢	0.11	-0.11 ¥ ⊙	0.09	<0.0001
8	Temporal	Temporal Pole	-0.04	0.11	-0.05	0.19	-0.20 ¢	0.17	-0.27 ¥ ⊙	0.29	<0.0001	-0.06	0.07	-0.09	0.11	-0.13	0.15	-0.18 ¥	0.14	<0.0001
9		Superior Temporal Gyrus, anterior division	-0.03	0.07	-0.04	0.11	-0.16 ¢ Ξ	0.10	-0.16 ¥ ⊙	0.11	<0.0001	-0.04	0.08	-0.06	0.10	-0.09	0.11	-0.12 ¥	0.10	0.0002
#		Superior Temporal Gyrus, posterior division	-0.04	0.08	-0.05	0.11	-0.16 ¢ Ξ	0.11	-0.16 ¥ ⊙	0.11	<0.0001	-0.04	0.08	-0.06	0.14	-0.08	0.10	-0.10	0.11	N.S.
#		Middle Temporal Gyrus, anterior division	-0.03	0.08	-0.06	0.13	-0.18 ¢ Ξ	0.13	-0.21 ¥ ⊙	0.15	<0.0001	-0.04	0.07	-0.07	0.14	-0.12 ¢	0.12	-0.10 ¥	0.10	<0.0001
#		Middle Temporal Gyrus, posterior division	-0.04	0.08	-0.06	0.12	-0.19 ¢ Ξ	0.15	-0.20 ¥ ⊙	0.12	<0.0001	-0.04	0.07	-0.06	0.12	-0.13 ¢	0.12	-0.14 ¥ ⊙	0.11	<0.0001
#		Middle Temporal Gyrus, temporo occipital part	-0.04	0.06	-0.04	0.10	-0.16 ¢ Ξ	0.12	-0.16 ¥ ⊙	0.11	<0.0001	-0.05	0.07	-0.06	0.12	-0.12	0.13	-0.11 ¥	0.10	0.0015
#		Inferior Temporal Gyrus, anterior division	-0.02	0.08	-0.07	0.13	-0.18 ¢	0.13	-0.21 ¥ ⊙	0.20	<0.0001	-0.04	0.06	-0.08	0.12	-0.13 ¢	0.12	-0.14 ¥	0.11	<0.0001
#		Inferior Temporal Gyrus, posterior division	-0.03	0.09	-0.07	0.13	-0.18 ¢ Ξ	0.15	-0.19 ¥ ⊙	0.18	<0.0001	-0.05	0.07	-0.08	0.10	-0.12 ¢	0.12	-0.14 ¥ ⊙	0.11	<0.0001
#		Inferior Temporal Gyrus, temporo occipital part	-0.02	0.09	-0.03	0.10	-0.15 ¢ Ξ	0.14	-0.16 ¥ ⊙	0.15	<0.0001	-0.05	0.08	-0.06	0.09	-0.11	0.12	-0.12 ¥	0.09	<0.0001
#		Parahippocampal Gyrus, anterior division	-0.05	0.13	-0.12	0.15	-0.18	0.16	-0.27 ¥ ⊙	0.27	<0.0001	-0.04	0.06	-0.08	0.08	-0.11 ¢	0.10	-0.14 ¥ ⊙	0.10	<0.0001
#		Parahippocampal Gyrus, posterior division	-0.03	0.16	-0.07	0.09	-0.11	0.14	-0.17 ¥ ⊙	0.21	0.0002	-0.02	0.05	-0.04	0.07	-0.08 ¢	0.08	-0.07 ¥	0.08	0.0001
#		Temporal Fusiform Cortex, anterior division	-0.02	0.15	-0.06	0.15	-0.15	0.12	-0.24 ¥ ⊙	0.30	<0.0001	-0.04	0.08	-0.07	0.09	-0.14 ¢	0.09	-0.17 ¥ ⊙	0.13	<0.0001
#		Temporal Fusiform Cortex, posterior division	-0.01	0.17	-0.03	0.11	-0.14	0.12	-0.20 ¥ ⊙	0.26	<0.0001	-0.05	0.08	-0.07	0.12	-0.15 ¢	0.11	-0.17 ¥	0.12	<0.0001
#		Heschl's Gyrus (includes H1 and H2)	-0.05	0.07	-0.06	0.11	-0.13 ¢	0.10	-0.15 ¥ ⊙	0.12	<0.0001	-0.05	0.11	-0.06	0.16	-0.09	0.11	-0.08	0.12	N.S.
#		Temporal Planum	-0.05	0.08	-0.05	0.11	-0.14 ¢ Ξ	0.10	-0.15 ¥ ⊙	0.12	<0.0001	-0.05	0.09	-0.04	0.12	-0.08 ¢	0.10	-0.08 ¥	0.11	N.S.

Table 4 - Longitudinal ROI-based analysis. Longitudinal average cortical thinning differences (mm), standard deviation (σ), and Tukey-Kramer multiple comparison post-hoc analysis in ANOVA (P). The data refer to: (d) CTR, (e) sMCI, (f) pMCI and (g) AD; α=0.01 level. ¢: Significant difference between CTR and pMCI; ¥: Significant difference between CTR and AD. Ξ: Significant difference between sMCI and pMCI; ⊙: Significant difference between sMCI and AD. N.S.: Not significant; CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's disease.

Civet detected significant longitudinal thinning rate changes also between sMCI and pMCI in 10 ROIs (check symbol Ξ) while Freesurfer could not find any variations. P values for multiple comparisons were always more significant in Civet ($P < 0.0001$).

Effect sizes

The effect sizes were derived as the Hedge's g (Figure 4). In the cross-sectional analysis, we decided to represent only CTR versus pMCI and versus AD, being these the combinations of highest interest when defining populations for disease-modifying and clinical trials. The effect size was always above 0.8 in those cortical regions expected to be heavily affected by the disease neuropathology. In CTR versus pMCI, Freesurfer's effect size was always higher. Only the posterior division of the temporal fusiform cortex was found to be statistically different ($p<0.05$) between the two pipelines. In CTR versus AD, the Hedge's g values followed the same trend for both algorithms without any statistical difference.

Longitudinally, Hedge's g trends were pretty similar for the two algorithms and increasing with the disease progression. No statistical differences were found in any ROIs or groups.

Cortical thickness versus cognitive impairment and hippocampal volumetry

Pearson's r correlation coefficients of regional cortical thickness with MMSE scores and quantitative hippocampal volume measurements (NeuroQuant − [26]) were investigated in each ROI (see Figure 5 panels A and B) within the CTR and pMCI patients, which represent the most appropriate population for innovative clinical trial designs.

In the CTR group, the relationship between pipelines' cortical thickness and cognitive function or hippocampal atrophy was generally weak ($-0.2 < r < 0.2$), cross-sectionally and longitudinally. This was expected due to the absence of the disease in these completely asymptomatic subjects. However, significant differences between Civet and Freesurfer were found in few areas (i.e.: frontal, parietal, occipital, and temporal).

In pMCI, the product momentums grew up to a medium and high levels ($-0.27 < r < 0.64$) especially for some expected ROIs, such as: precuneus cortex, cingulate and parahippocampal gyri. Significant differences between Civet and Freesurfer were found in a number of ROIs (i.e.: frontal, parietal, occipital, limbic, and temporal). Both Civet and Freesurfer cortical thickness measurements correlate better with hippocampal atrophy measurements than with neuropsychological tests.

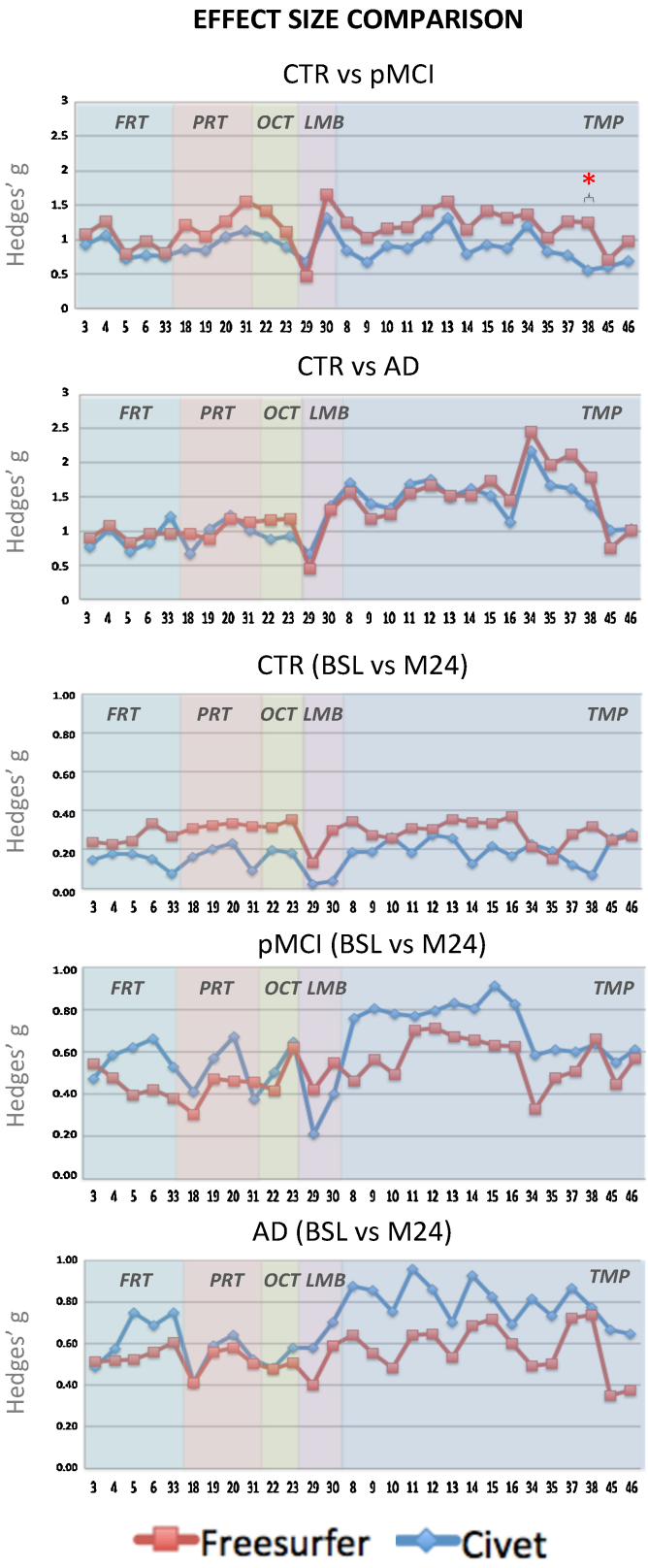


Figure 4 - Hedges' g effect size graphs in the different ROI areas: The first two panels represent the cross-sectional effect sizes comparing the overall trend of CTR versus pMCI, and of CTR versus AD. The remaining three panels represent the longitudinal effect sizes between the baseline and month 24 in CTR, pMCI, and AD groups. The * symbol stands for $p<0.05$.

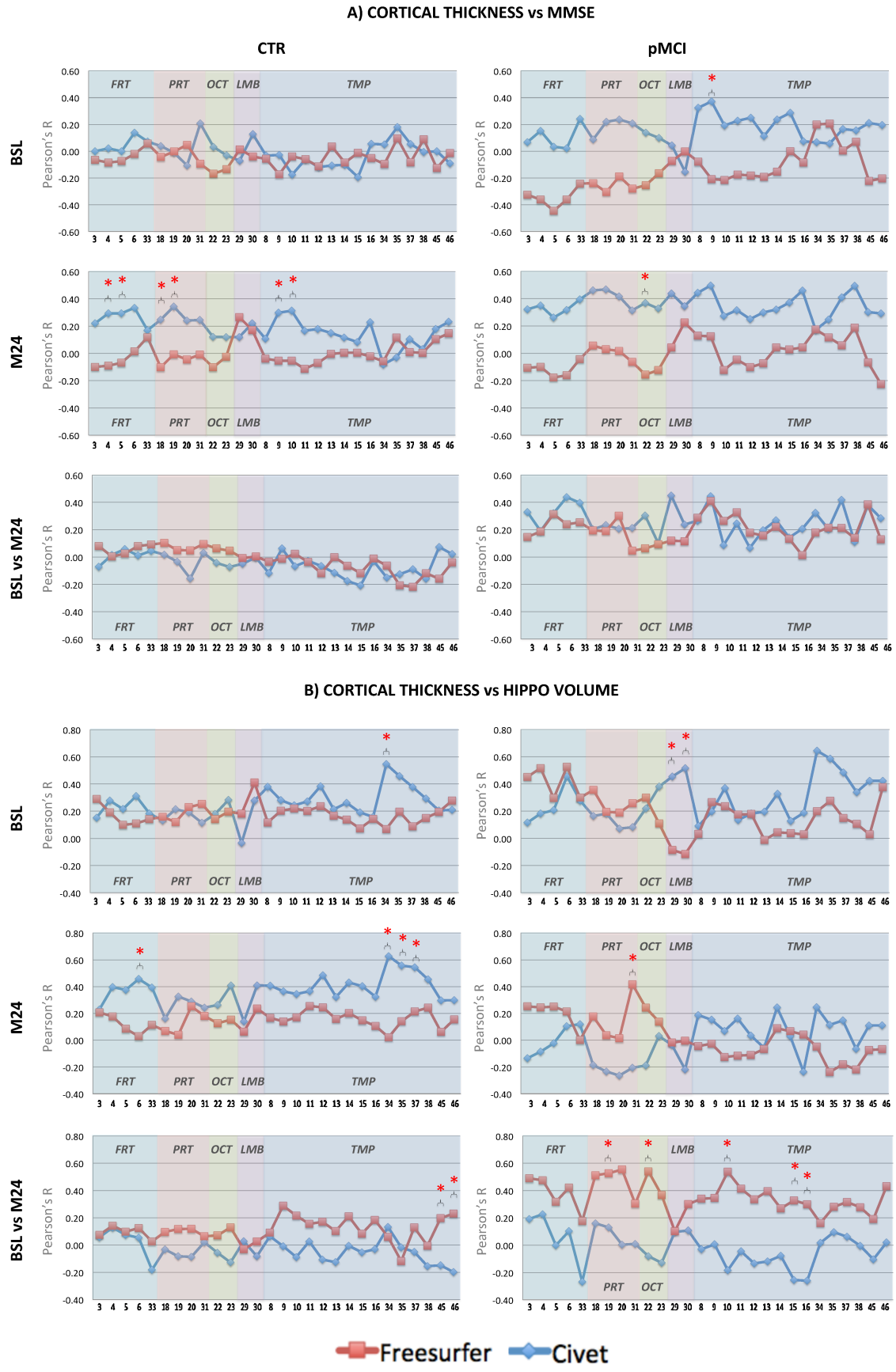


Figure 5 - Pearson's r coefficient of cortical thickness versus MMSE scores (**panel A**): In the CTR group, no significant differences between ROIs were detected in the two pipelines at BSL. At M24, significant differences

between the two pipelines were found in the: middle frontal gyrus; inferior frontal gyrus - pars triangularis; superior parietal lobule; anterior division of the supramarginal gyrus; anterior and posterior division of the superior temporal gyrus. Longitudinally, no significant differences between ROIs were detected in the two pipelines. In the pMCI group, significant difference between the two pipelines was found at BSL in the: anterior division of the superior temporal gyrus. At M24, significant difference between the two pipelines was found in the: superior division of the lateral occipital cortex. Longitudinally, no significant differences between ROIs were detected in the two pipelines.

Pearson's r coefficient of cortical thickness versus NeuroQuant® hippocampal volume (**panel B**): In the CTR group, significant difference between the two pipelines at BSL was found in the: anterior division of the parahippocampal gyrus. At M24, significant differences between the two pipelines were found in the: inferior frontal gyrus - pars opercularis; anterior and posterior division of the parahippocampal gyrus; anterior division of the temporal fusiform cortex. Longitudinally, significant differences between the two pipelines were found in the: Heschl's gyrus and temporal planum. In the pMCI group, significant difference between the two pipelines was found at BSL in the: precuneus cortex. Longitudinally, significant differences between the two pipelines were found in the: anterior division of the supramarginal gyrus, superior division of the lateral occipital cortex, posterior division of the superior temporal gyrus, posterior division of the inferior temporal gyrus, temporo-occipital part of the inferior temporal gyrus. In panels A and B, * symbol stands for $p<0.05$ (Steiger's z-test). Red coloured lines represent the trends in Freesurfer, blue lines in Civet. CTR.: CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's disease; BSL: baseline; M24: month 24; FRT: Frontal; PRT: Parietal; OCT: Occipital; LIMB: Limbic; TMP: Temporal.

ROC Analysis

Figure 6 shows the Receiver Operating Characteristic (ROC) curves used to discriminate pMCI and AD patients from the CTR group at baseline, together with the longitudinal cortical pattern used to discriminate pMCI. Identifying the most informative ROI was mandatory to reduce the dimensionality problem. In order to maximize the discriminatory power, we adopted a sequential forward search strategy (i.e., adding successive ROIs to the target set) as feature selection criterion. The goal was to find the best combination of ROIs for both tools with the highest discriminatory power. The best ROIs used to generate the final ROCs were different in each curve and for each algorithm. We started selecting those ROI with the highest effect size; at each further step, we assessed other ROIs with a medium-large effect size ($d > 0.8$ in cross sectional analysis; $d > 0.6$ in longitudinal analysis). This process reduced the inherent noise of high-resolution data, as well as the risk of over-fitting. Logistic regressions on regional cortical thickness in the selected combinations of ROIs were performed to build ROC curves, AUCs and the relative Intervals of Confidence (CI). No statistical difference ($p>0.05$) was found between the AUCs derived with Civet and those derived with Freesurfer. At baseline, CTR versus pMCI yielded 0.8953 and 0.9313 ($z=-0.46$, $r=0.31$), while CTR versus AD yielded

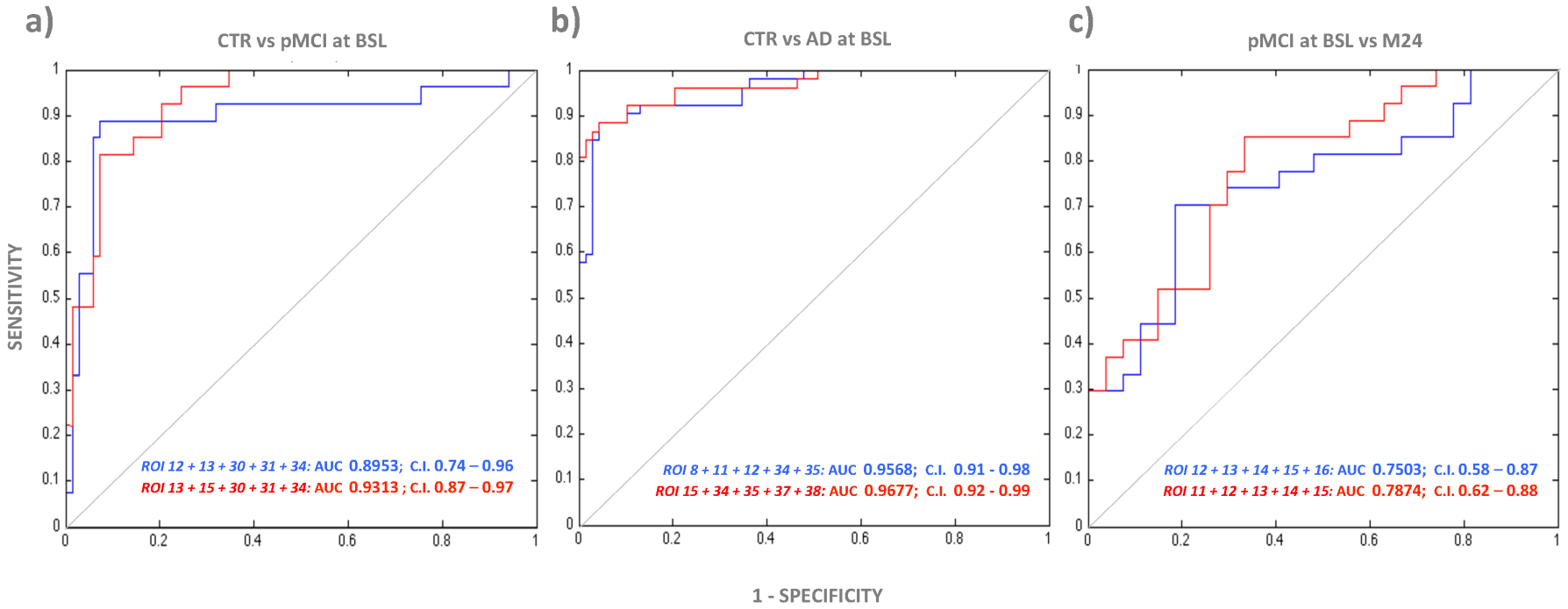


Figure 6 - Receiving Operator Characteristic (ROC) curves showing the performances of Civet and Freesurfer in classifying: A) CTR versus pMCI at baseline; B) CTR versus AD at baseline; and C) pMCI at baseline from month 24. AUC with 95% CIs are reported for both Freesurfer in red and Civet in blue. CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's Disease; BSL: baseline; M24: month 24; AUC: Area Under the Curve; C.I: Confidence Interval; ROI 8: temporal pole; ROI 11: anterior division of the middle temporal gyrus; ROI 12: posterior division of the middle temporal gyrus; ROI 13: temporo-occipital part of middle temporal gyrus; ROI 15: posterior division of inferior temporal gyrus; ROI 16: temporo-occipital part of inferior temporal gyrus; ROI 30: posterior division of the cingulate gyrus; ROI 31: Precuneus Cortex; ROI 34: anterior division of the parahippocampal gyrus; ROI 35: posterior division of the parahippocampal gyrus; ROI 37: anterior division of the temporal fusiform cortex; ROI 38: posterior division of the temporal fusiform cortex.

0.9568 and 0.9677 respectively (z=-0.38, r=0.46). In the longitudinal framework, pMCI yielded 0.7503 and 0.7874 (z=-0.34, r=0.21). Freesurfer performed slightly better in terms of classification accuracy, both on cross sectional and longitudinal analyses.

DISCUSSION

This study could be considered as a first attempt to verify the mutual strengths and weaknesses of Civet and Freesurfer in a real head-to-head challenge, at the precision level of the single voxel. In the literature, only phantom-based validation methods have been used [27,28] but this kind of approach does not take into consideration every aspects of real data. We investigated and compared the performances of Civet and Freesurfer when applied to the same ADNI1 groups which included subjects on the entire disease spectrum, as monitored in a 2-year time frame. The analyses showed commonalities and differences.

Civet and Freesurfer are characterized by specific and distinctive procedures, making it difficult to compare their outputs. This problem was solved adopting a combined approach, applying both the GVF and CPS to ensure a robust comparison of meshes characterized by different morphometry and topography completely different. Thanks to the direct vertex-by-vertex cross-algorithm comparison, the differences between the two algorithms, with regard both to cross-sectional and longitudinal analysis, were analytically mapped.

Differences between thickness evaluation of the first test (MPRAGE) and that of the retest (MPRAGE-Repeat) did not appear, suggesting high repeatability. Both Civet's and Freesurfer's performances changed according to the disease stage, pointing out that neither algorithm can be considered better than the other, or the best acting. Freesurfer systematically underestimated the absolute thickness by about 1 mm if compared to Civet's performance. Explanations for this evidence are not trivial. However, the restriction of Freesurfer to 1.0 mm as resolution for the volumes to be processed could be one possible reason. Civet, relying on the volumetric Laplacian approach, can use higher resolutions (e.g.: 0.8 or 0.9 mm) often adopted in ADNI1. An important role might be also played by the different mathematical procedures used by the two tools when reconstructing the gray matter sheet. Moreover, the skeleton reconstruction method adopted by Civet to build the GM sheet tends to overestimate the cortical thickness in case of blurred regions (i.e.: regions affected by noise where CSF volume is small); on the other hand, Freesurfer relies on the inner white deformation surface approach, which can be strongly influenced by the anatomical accuracy of the surface reconstruction at both inner and outer boundaries, thus giving a partially unfair anatomical accuracy of the surface reconstruction and assessment of the cortical thickness.

Cross-sectionally, both algorithms were sensitive to cortical thinning in those cortical regions heavily affected by the neuropathology. Comparing CTR to pMCI, the regions of significance found by both tool were overlapping with the those found comparing CTR and AD, albeit

smaller, indicating that the differences in cortical thinning are progressive and well detectable even before a formal diagnosis of AD. This means that both tools can detect the characteristic signature of AD. Both Civet and Freesurfer were able to efficiently differentiate CTR from the AD and pMCI. All the ROIs granting such a good discrimination rate belonged to the temporal lobe. An interesting consideration for future works is the possibility to use Civet and Freesurfer to differentiate AD in particular subclasses, namely familial AD, early onset AD, and late onset AD [29,30].

Longitudinally, both pipelines showed more statistically atrophic clusters in CTR than in sMCI, but this should be considered as a confounding phenotypic effect due to demographic, numerosity, clinical and other genetic characteristics. Further analyses with a larger sample will be conducted to clarify this particular behaviour. In pMCI, Civet was able to highlight a characteristic atrophic pattern involving expected temporal areas, such as the inferior margin of central gyrus and extended lateral frontal-parietal areas, as expected. The Civet's higher effect size and its more representative cortical signature suggest that this tool can detect the typical atrophic patterns in subject that will convert to AD within 2 years more efficiently. In the discriminant analysis, Civet produced an AUC slightly lower than that produced by Freesurfer; but this was probably due to random noises that confuses classifiers, producing changes hard to predict and control. Additional explanation can be related to the fact that longitudinally, on a vertex-by-vertex basis, Civet showed a more extensive effect than Freesurfer, while on a ROI basis the differences between the pipelines were not significant. In the AD cohort both Freesurfer and Civet were analogously sensitive to the thinning patterns. As far as the correlation between the cortical thinning and hippocampal atrophic rate is concerned, Freesurfer showed a better trend, probably due to the exploitation of the longitudinal stream.

Given its progressive alteration along the MCI-to-AD course, cortical thickness seems to be a promising neuroimaging candidate marker. With few exceptions, the two algorithms showed robust multi-ROI correlation patterns fairly consistent with the usual clinical and regional neuroimaging biomarkers, thus producing new, 3D, global profiles of the disease progression. Ultimately, having reliable 3D diagnostic markers would enable clinicians to identify and treat MCI patients who will evolve into AD patients in a timely manner, as disease-modifying treatments will become available.

Future studies, including the MR 3.0 Tesla field strength, additional time points, extended age range of subject, larger and additional groups, might be helpful to further address the spatial and temporal atrophic pattern of the Alzheimer's changes.

Freesurfer and Civet have been validated against either histological analysis or manual measurements [31-34], but none of them has been contrasted against different stages of the Alzheimer's pathology. Future works should focus on further validating both pipelines against a database of cortical thickness derived from a population of normal and abnormal cadaveric brains, such as those recently defined in the BigBrain initiative (https://bigbrain.loris.ca/).

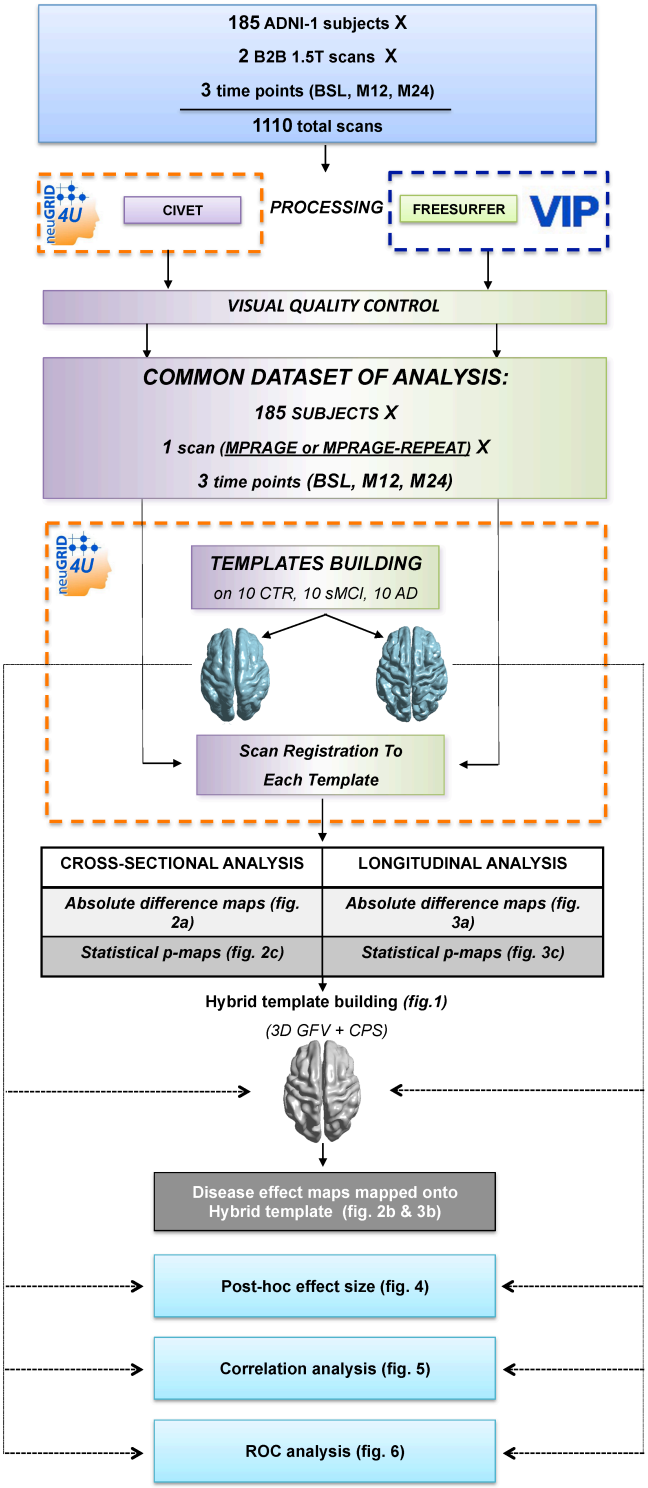
Some limitations should be considered in the interpretation of the present results. First, the tools here described need to be further compared with other recent available techniques, such as: Toads-Cruise [35], ARCTIC [36], MILXCTE [37], DiReCT [38], or CLADA [39]. Second, as expert manual rater in neuroimaging represents the gold standard, independent evaluators should compare the performance and accuracy of each automatic pipeline. Third, each tool should be validated against harmonized MR datasets, such as: standardized ADNI analysis dataset [40] WW-ADNI [41], AddNeuroMed [42] and OASIS [43]. Fourth, computational time is worth consideration: the extensive use of Civet or Freesurfer to analyse large volumes of data mandatorily requires HPC, Grid or Cloud resources, due to the protracted processing time needed. Additional developing and programming can make these algorithms more reliable, faster and slighter.

CONCLUSION

Both Civet and Freesurfer demonstrated high sensitivity to cortical gray matter changes cross-sectionally and longitudinally. Additional efforts are needed to clarify the ability of these tools to address particular clinical and research questions concerning the future use of cortical thickness as a biomarker, and in particular their ability to: (I) predict cortical decline along different time points, (II) reduce the number of patients needed for future clinical trials, (III) help monitoring the efficacy of disease modifying drugs.

SUPPLEMENTAL DATA

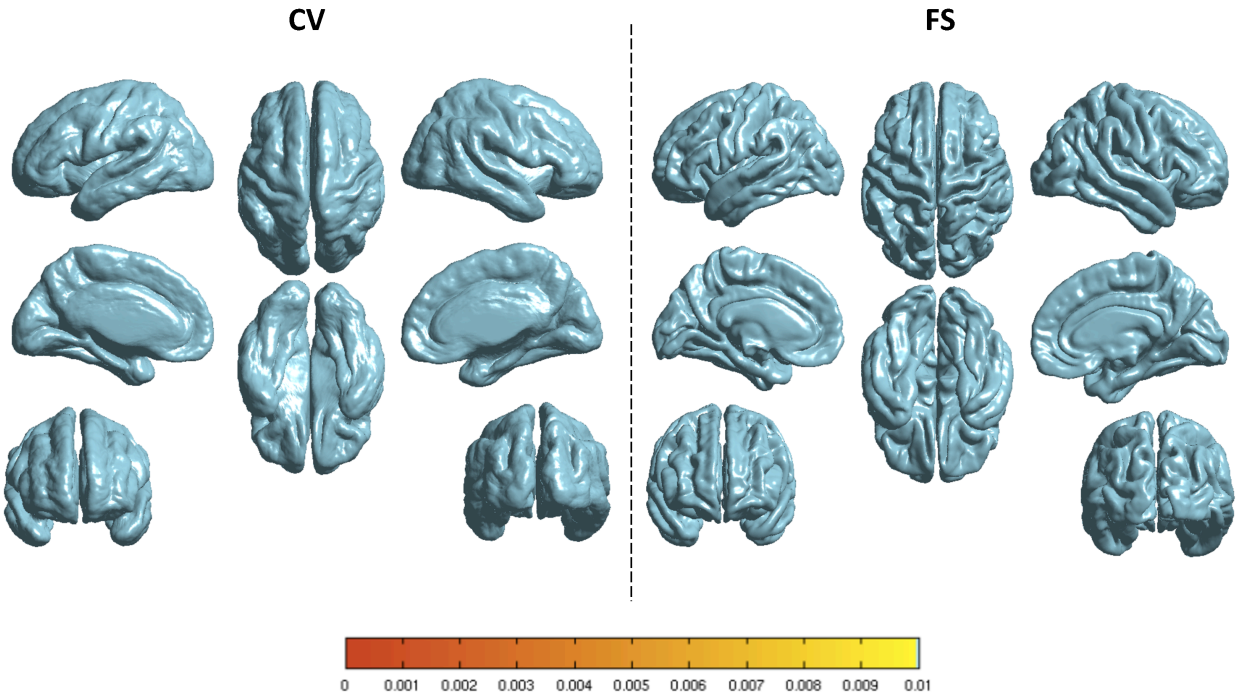
S1 Figure. Flowchart of the study methodology.



Two pipelines for the extraction of the cortical thickness (CIVET-CLASP v1.1.9 & FREESUERFER v5.3.0) over 1110 ADNI1 back to back (B2B; i.e.: MPRAGE and MPRAGE-Repeat) scans were compared. The dataset of analysis used was composed by: 69 CTR; 37 sMCI; 27 pMCI and 52 AD. For each subject three time points (i.e.: baseline, month 12 and month 24) were considered. Images were processed using two interconnected e-infrastructures: the neuGRID (http://neugrid4you.eu) and VIP (http://vip.creatis.insa-lyon.fr) platforms. At the end of the computing tasks, an expert evaluator performed a visual quality control on the 3D cortical meshes generated by the algorithms. For visualization purposes, a reference template was created for each pipeline by averaging the same 10 CTR, 10 sMCI and 10 AD brain surfaces. Then, starting from those, the hybrid template was generated making use of the 3D gradient vector flow (GVF) and closest point search (CPS) techniques. This enabled a direct comparison between Civet and Freesurfer results, minimizing the morphological differences and making the vertex-by-vertex comparison possible. 28 Regions Of Interest (ROIs) were then defined to capture the spatially distributed pattern of cortex atrophy most affected by Alzheimer's disease. To investigate the consistency of cortical thinning measurements, we extracted the thickness from the ROIs of interest in each study group. Cross sectional and longitudinal thinning maps were generated for each group in term of absolute cortical difference, disease effects, and statistical significance estimates. The effect size analysis of the regional cortical thinning was derived through the post-hoc Hedgen's g index. Correlation analysis between ROI thickness, Mini Mental State Examination (MMSE) scores, as well as the hippocampal volumetry (computed with NeuroQuant®) were performed. Receiver Operating Characteristic (ROC) curves and the relative Area Under the Curve (AUC) with the Confidence Interval (CI) represented the performance of a binary classifier system for discriminative analyses relying on Civet's and Freesurfer's cortical thickness measures. BSL: Baseline; M12: Month 12; M24: Month 24; CV: Civet; FS: Freesurfer. The following is a brief description of the atomic elements composing the study workflow as a whole:

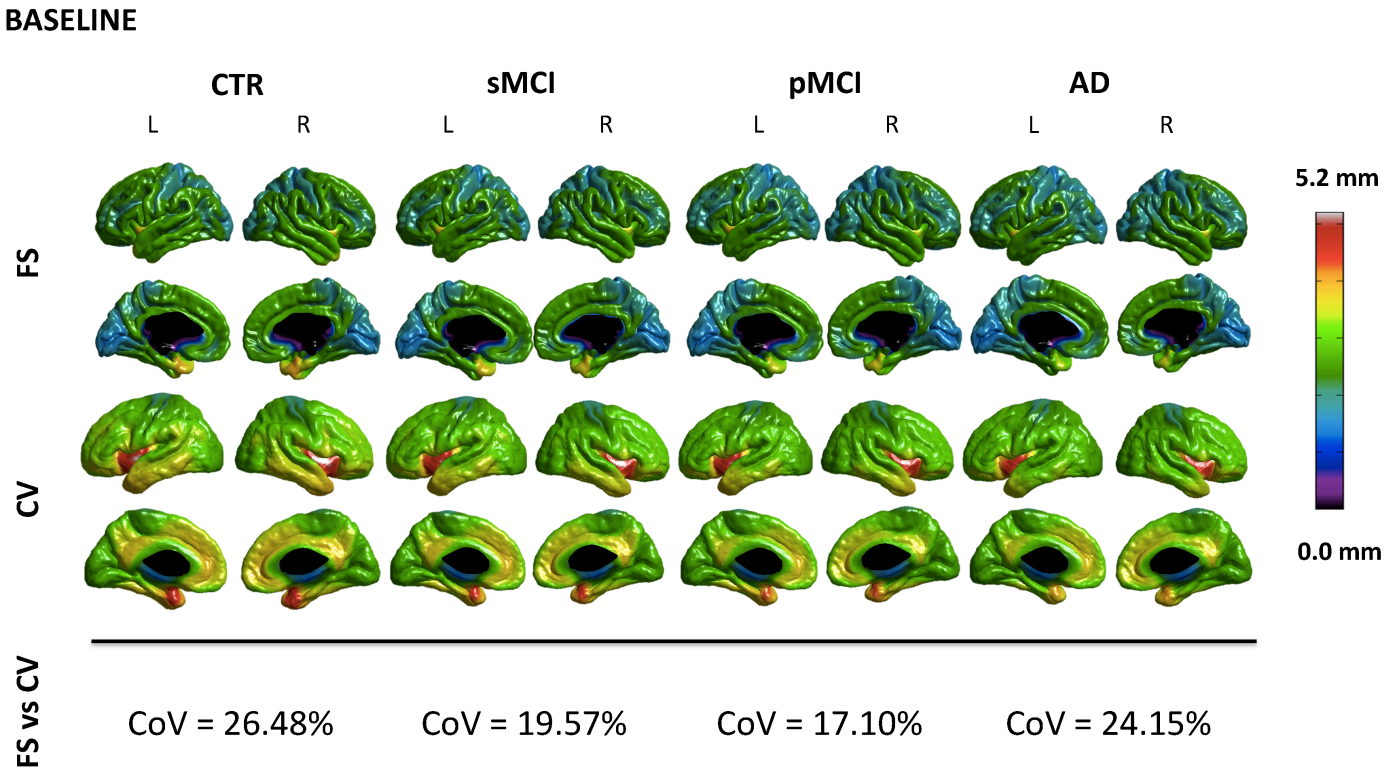
- Blue box denotes the starting number of image collected for the present study
- Data analyzed by the CIVET pipeline (v.1.1.9)
- Data analyzed by FREESURFER pipeline (v.5.3.0)
- Common data processed by both algorithms
- Data processed in N4U
- Data processed in VIP
- Absolute difference maps in mm
- Statistical maps which highlight regions with p<0.01
- Maps generated using the hybrid template
- Turquoise boxes denote statistical analyses (e.g.: Hedges' g, Pearson R, ROC)

S2 Figure. Civet and Freesurfer B2B repeatability.



Both algorithms showed excellent repeatability. No significant differences between B2B (MPRAGE and MPRAGE-repeat) acquisitions are found (FDR corrected p-maps ($p < 0.01$)). The maps are generated from the common dataset used in this study (69 CTR, 37 sMCI, 27 pMCI and 52 AD). Acronyms: CV.: Civet template; FS.: Freesurfer template; L.: Left hemisphere; R.: Right hemisphere; B2B: ADNI back to back volumetric acquisition scans (MPRAGE and MPRAGE Repeat). CTR.: Normal Elderly Controls; sMCI.: stable MCI; pMCI.: progressive MCI; AD.: Alzheimer's disease.

S3 Figure. Freesurfer and Civet absolute cortical thickness maps for every diagnostic class.



Civet shows 1 mm greater thickness than Freesurfer. At naked eyes, no harsh variations over the crests of gyri and depths of sulci are visible. CoV of the CV and FS difference is reported per each diagnostic class. The higher the CV, the bigger the residuals relative to the predicted value of the two pipelines. CV: Civet; FS: Freesurfer; CoV: Coefficient of variation; CTR: Normal elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's Disease

S1 Table. Link to the list of subjects' RIDs:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4364123/bin/pone.0117692.s004.xlsx

S2 Table. Whole brain absolute mean cortical thickness (mm) ± standard deviation (σ) for each diagnostic group at baseline and month 24.

		CIVET			FREESURFER		
	N	**Baseline**	**Month 24**	**^{I}P**	**Baseline**	**Month 24**	**^{I}P**
CTR	69	3.18 ± 0.13	3.15 ± 0.13	0.0005	2.19 ± 0.11	2.14 ± 0.12	<0.0001
sMCI	37	3.09§ ± 0.14	3.06 ± 0.13	N.S.	2.11 ± 0.12	2.06 ± 0.15	0.0015
pMCI	27	3.04$^{¢}$ ± 0.13	2.94$^{¢,Ξ}$ ± 0.16	0.0001	2.02$^{¢}$ ± 0.13	1.94$^{¢,Ξ}$ ± 0.13	<0.0001
AD	52	2.99$^{¥,⊙}$ ± 0.15	2.85$^{¥,⊙}$ ± 0.17	<0.0001	2.00$^{¥,⊙}$ ± 0.14	1.90$^{¥,⊙}$ ± 0.16	<0.0001
^{II}P		<0.0001	<0.0001		<0.0001	<0.0001	

The cross-sectional comparison revealed a statistically significant variation at baseline and month 24 (p <0.0001) for both pipelines. At baseline, Civet was much more sensitive than Freesurfer, as shown by the ability of the former to detect a statistical difference between CTR and sMCI. Both pipelines at month 24 were able to detect significant differences between sMCI and pMCI. Acronyms: N.S.: not significant; ^{I}P: paired samples t-test; ^{II}P: Tukey-Kramer multiple comparison post hoc analysis in one-way ANOVA. α=0.01 level. §: Significant difference between CTR and sMCI; ¢: Significant difference between CTR and pMCI; ¥: Significant difference between CTR and AD; Ξ: Significant difference between sMCI and pMCI; ⊙: Significant difference between sMCI and AD. CTR: Healthy elderly controls; sMCI: stable MCI; pMCI: progressive MCI; AD: Alzheimer's Disease.

S3 Table. Cross-sectional thinning percentages (%) ± standard deviation (σ) in paired diagnostic groups at baseline.

	CROSS SECTIONAL THINNING PERCENTAGE (Baseline)	
	CIVET	**FREESURFER**
CTR vs sMCI	2.63 ± 6.13	3.42 ± 7.43
CTR vs pMCI	4.41 ± 5.79	7.71 ± 7.79
CTR vs AD	6.03 ± 6.25	8.47 ± 8.50
P	N.S.	N.S.

Civet was characterized by a pretty constant thinning rate along with the progression of the disease; while Freesurfer seemed to peak earlier, reaching its steepest shrinkage when comparing CTR versus pMCI. Significance (P) was tested using Tukey-Kramer multiple comparison post hoc analysis in one-way ANOVA (α=0.05 level).

S4 Table. Longitudinal thinning percentage (%) ± standard deviation (σ) in each diagnostic group in a time span of two years.

LONGITUDINAL THINNING PERCENTAGE (Baseline to Month 24)		
	CIVET	**FREESURFER**
CTR	0.78 ± 1.78	2.02 ± 2.56
sMCI	1.03 ± 2.68	2.28 ± 4.04
pMCI	$3.28^{¢}$ ± 3.01	4.09 ± 3.87
AD	$4.50^{¥,⊙}$ ± 3.62	$4.96^{¥,⊙}$ ± 3.71
P	<0.0001	0.0001

The significance was tested using Tukey-Kramer multiple comparison post hoc analysis in one-way ANOVA (α=0.01 level). ¢: Significant difference between CTR and pMCI; ¥: Significant difference between CTR and AD; ⊙: Significant difference between sMCI and AD. CTR.: Normal Elderly Controls. sMCI.: stable MCI. pMCI.: progressive MCI. AD.: Alzheimer's disease.

DATA AVAILABILITY

The data generated in this study are made publicly available to promote the evaluation of cortical thickness tool. Data may be accessed as "H2H Comparison Study" at https://neugrid4you.eu/datasets.

REFERENCES

1. Fox NC, Kennedy J (2009), Structural imaging markers for therapeutic trials in Alzheimer's disease. J Nutr 13(4): 350-352.
2. Jack CR Jr, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, et al. (2004), Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 62(4): 591-600.
3. Fox NC, Black RS, Gilman S, Rossor MN, Griffith SG, et al. (2005), Effects of Abeta immunization (AN1792) on MRI measures of cerebral volume in Alzheimer disease. Neurology 64(9): 1563-1572.
4. Thompson P, Hayashi KM, Zubicaray G, Janke AL, Rose SE, et al. (2003), Dynamics of gray matter loss in Alzheimer's disease. J Neurosci 23: 994–1005.
5. Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, et al. (2006), Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage 32(1): 180-94.
6. Dickerson BC, Sperling RA (2005), Neuroimaging biomarkers for clinical trials of disease-modifying therapies in Alzheimer's disease. NeuroRx 2: 348–360.
7. MacDonald D, Kabani N, Avis D, Evans AC (2000), Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12(3): 340-356.
8. Fischl B and Dale AM (2000), Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci U S A 97(20): 11050-11055.
9. Dahnke R, Yotter RA, Gaser C (2013), Cortical thickness and central surface estimation. Neuroimage 65:336-48.
10. Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab'bagh Y, et al. (2005), Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a laplacian map and partial volume effect classification. Neuroimage 27(1): 210-221.
11. Dale AM, Fischl B, Sereno MI (1999), Cortical surface-based analysis. I. segmentation and surface reconstruction. Neuroimage 9(2): 179-194.
12. Frisoni GB, Fox NC, Jack CR Jr, Scheltens P, Thompson PM (2010), The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6(2): 67-77.
13. Clarkson MJ, Cardoso MJ, Ridgway GR, Modat M, Leung KK, et al. (2011), A comparison of voxel and surface based cortical thickness estimation methods. Neuroimage 57(3): 856-865.
14. Redolfi A, Bosco P, Manset D, Frisoni GB; neuGRID consortium (2013), Brain investigation and brain conceptualization. Funct Neurol 28(3): 175-190.
15. Glatard T, Lartizien C, Gibaud B, da Silva RF, Forestier G, et al. (2013), A virtual imaging platform for multi-modality medical image simulation. IEEE Trans Med Imaging 32(1): 110-118.Reuter M, Schmansky NJ, Rosas HD, Fischl B (2012), Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61(4): 1402-1418.
16. Desikan RS, Segonne F, Fischl B, Quinn BT, Dickerson BC, et al. (2006), An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31(3): 968-980.
17. Dickerson BC, Feczko E, Augustinack JC, Pacheco J, Morris JC, et al. (2009a), Differential effects of aging and alzheimer's disease on medial temporal lobe cortical thickness and surface area. Neurobiol Aging 30(3): 432-440.
18. Dickerson BC, Bakkour A, Salat DH, Feczko E, Pacheco J, et al. (2009b), The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb Cortex 19(3): 497-510.
19. Liu T, Nie J, Tarokh A, Guo L, Wong ST (2008), Reconstruction of central cortical surface from brain MRI images: Method and application. Neuroimage 40(3): 991-1002.

20. Querbes O, Aubry F, Pariente J, Lotterie JA, Démonet JF, et al. (2009), Early diagnosis of Alzheimer's disease using cortical thickness: Impact of cognitive reserve. Brain 132: 2036-2047.
21. Genovese CR, Lazar NA, Nichols T (2002), Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15(4): 870-878.
22. Hanley JA and McNeil BJ (1983), A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148(3): 839-843.
23. Lerch JP and Evans AC (2005), Cortical thickness analysis examined through power analysis and a population simulation. Neuroimage 24(1): 163-173.
24. Jones R, Payne R (1997) Clinical Investigation and Statistics in Laboratory Medicine (Management & Technology in Laboratory Medicine). ACB Venture Publications. 188 p. ISBN-10: 0902429213. ISBN-13: 978-0902429215.
25. Brewer JB, Magda S, Airriess C, Smith ME (2009), Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer Disease. Am J Neuroradiol 30(3): 578-580.
26. Lee JK, Lee JM, Kim JS, Kim IY, Evans AC, et al. (2006a), A novel quantitative cross-validation of different cortical surface reconstruction algorithms using MRI phantom. Neuroimage 31(2): 572-584.
27. Lee J, Lee JM, Kim JH, Kim IY, Evans AC, et al. (2006b), A novel quantitative validation of the cortical surface reconstruction algorithm using MRI phantom: issues on local geometric accuracy and cortical thickness. Med Image Comput Comput Assist Interv 9: 183-190.
28. Knight WD, Kim LG, Douiri A, Frost C, Rossor MN, et al. (2011), Acceleration of cortical thinning in familial Alzheimer's disease. Neurobiol Aging 32(10): 1765-1773.
29. Ridgway GR, Lehmann M, Barnes J, Rohrer JD, Warren JD, et al. (2012), Early-onset Alzheimer disease clinical variants: multivariate analyses of cortical thickness. Neurology 79(1): 80-84.
30. Kabani N, Le Goualher G, MacDonald D, Evans AC (2001), Measurement of cortical thickness using an automated 3-D algorithm: A validation study. Neuroimage 13(2): 375-380.
31. Kuperberg GR, Broome MR, McGuire PK, David AS, Eddy M, et al. (2003), Regionally localized thinning of the cerebral cortex in schizophrenia. Arch Gen Psychiatry 60(9): 878-888.
32. Rosas HD, Liu AK, Hersch S, Glessner M, Ferrante RJ, et al. (2002), Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 58(5): 695-701.
33. Salat DH, Buckner RL, Snyder AZ, Greve DN, Desikan RS, et al. (2004), Thinning of the cerebral cortex in aging. Cereb Cortex 14(7): 721-730.
34. Han X, Pham DL, Tosun D, Rettmann ME, Xu C, et al. (2004), CRUISE: Cortical reconstruction using implicit surface evolution. Neuroimage 23(3): 997-1012.
35. Hazlett HC, Poe MD, Gerig G, Styner M, Chappell C, et al. (2011), Early brain overgrowth in autism associated with an increase in cortical surface area before age 2 years. Arch Gen Psychiatry 68(5): 467-476.
36. Acosta O, Fripp J, Doré V, Bourgeat P, Favreau JM, et al. (2012), Cortical surface mapping using topology correction, partial flattening and 3D shape context-based non-rigid registration for use in quantifying atrophy in Alzheimer's disease. J Neurosci Methods 205(1): 96-109.
37. Das SR, Avants BB, Grossman M, Gee JC (2009), Registration based cortical thickness measurement. Neuroimage 45: 867–879.
38. Nakamura K, Fox R, Fisher E (2011), CLADA: Cortical longitudinal atrophy detection algorithm. NeuroImage 54(1): 278-289.

39. Wyman BT, Harvey DJ, Crawford K, Bernstein MA, Carmichael O, et al. (2013), Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimers Dement 9(3):332-7.
40. Carrillo MC, Bain LJ, Frisoni GB, Weiner MW (2012), Worldwide Alzheimer's disease neuroimaging initiative. Alzheimers Dement 8(4): 337-342.
41. Westman E, Simmons A, Muehlboeck JS, Mecocci P, Vellas B, et al. (2011), AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. Neuroimage 58(3): 818-828.
42. Ardekani BA, Figarsky K, Sidtis JJ (2013), Sexual Dimorphism in the Human Corpus Callosum: An MRI Study Using the OASIS Brain Database. Cereb Cortex 23(10): 2514-2520.

Chapter 4

QCE (Quality Control Environment): A machine learning tool for automatic classification of cortical meshes

Alberto Redolfi
Clara Fischer
Daniele Orlandi
Édouard Duchesnay
Grégory Operto
Denis Rivière
Giovanni B Frisoni
Jean-François Mangin
NeuGRID and CATI consortia
Alzheimer's Disease Neuroimaging Initiative

NeuroImage
Submitted

ABSTRACT

Human neuroimaging has entered the big-data era. Neuroscientists can process their data more and more efficiently but one major barrier remains: the manual validation of automatic processing results.

To address this issue, an automatic Quality Control Environment (QCE) was designed. It is a machine learning (ML) multi-label classifier for predicting 3D cortical mesh artefacts of two widely used pipelines: Freesurfer (FS) and Civet-CLASP (CV).

QCE was validated against the gold-standard visual assessment of experts. Cortical meshes (FS=1'582; CV=1'692) from five datasets (ADNI, ARWIBO, EDSD, OASIS, PharmaCOG) were considered. QCE is a multi-level classifier. In Level-1, a Random Forest (RF) estimator allows binary classification of "good" versus "bad" cortical meshes. In Level-2, a soft ensemble classifier composed of a Support Vector Machine (SVM), plus a RF allows the multi-label tagging of eight artefacts (widespread, temporal, insula, parietal, frontal, meninges, occipital, other problems) on "bad" meshes.

In Level-1, QCE appeared to be slightly more accurate, sensitive and with lower type II errors in FS than CV. On the contrary, QCE was more specific and had lower type I errors in CV than FS. The QCE area under the curve (AUC) yielded good results (AUC^{FS}_{Left}= 0.96; AUC^{FS}_{Right}=0.94; AUC^{CV}_{Left}=0.96; AUC^{CV}_{Right}=0.91) although significant differences were measured in left versus right hemispheres ($p<0.05$) in both pipelines.

In Level-2, the Hamming-Loss (HL) of the QCE was systematically lower in FS (HL^{FS}_{Left}= 0.19; HL^{FS}_{Right}= 0.21; HL^{CV}_{Left}= 0.22; HL^{CV}_{Right}= 0.23). The Label Ranking Average Precision (LRAP) score was always lower in CV ($LRAP^{FS}_{Left}$= 0.62; $LRAP^{FS}_{Right}$= 0.59; $LRAP^{CV}_{Left}$= 0.51; $LRAP^{CV}_{Right}$= 0.51). HL and LRAP metrics suggested QCE can well-perform to detect multiple artefacts in both pipelines.

QCE appears to be a robust, reliable and fully automatic system capable of generalizing from training examples and correctly classifying any new 3D cortical mesh. QCE offers the neuroscientists an opportunity to quality control data more systematically, spending their time and resources entirely on data analysis.

INTRODUCTION

Quality control (QC) of 3D cortical meshes from current analysis pipelines is a difficult task due to the possibility of numerous types of errors. No algorithm or pipeline yet available today in neuroimaging can generate outputs which are 100% correct. QC and output validation procedures can only be carried out by expert scientists as a visual inspection [17, 9].

For most exploited algorithms in cortex delineations, such as Freesurfer [10] and Civet-CLASP [15], some semi-automatic QC procedures to detect outliers and collect detailed snapshots of various brain cortical regions started to appear over the last few years.

These are namely:

(i) QA-TOOL[1]: its aim is to highlight subjects with potential 3D mesh problems. It is not considered to be a substitute for manual inspection of each subject slice, but it does allow for a quick scan of the processed output to discover any obvious failures. This tool has been developed specifically for the Freesurfer pipeline.

(ii) CBRAIN-QC[2]: this generates a summary report containing group statistics and is linked to individual screenshots. A table reporting various metrics and colour codes per mesh indicate any possible delineation warnings or errors. It was developed specifically for the Civet pipeline.

(iii) ENIGMA [28]: this identifies subjects with cortical values (i.e., thickness, surface or volume) that deviate from the specific population being studied. It creates cortical surface snapshots from internal brain slices, as well as external 3D views from different angles. It is available for both the Freesurfer and Civet pipelines. However, all these tools still need direct control by an expert.

The main problems affecting 3D cortical mesh can be categorized into eight categories, which are illustrated in Figure 1. There may be several errors within a cortical mesh and normally these are very difficult to detect, requiring subtle attention.

As in many other research fields, neuro-imaging is also experiencing an introduction of big-data. This means particularly, an exponential growth of open-access brain imaging data as well as fast surrogate imaging biomarkers availability through more and more common grid/cloud e-infrastructures [12, 23, 26, 3]. Therefore, the traditional means to visually inspect results is no longer sustainable if we consider that thousands of brain images can be quickly processed in a high through-put way.

[1] https://surfer.nmr.mgh.harvard.edu/fswiki/QATools
[2] http://www.bic.mni.mcgill.ca/ServicesSoftware/QualityControl

To face this change of paradigm in neuro-imaging, crowdsourcing for error detection and large scale data annotation has recently been proposed [14]. This approach is based on outsourcing laborious QC tasks to anonymous workers from an online community. Although this is feasible, additional effort is needed to increase the sensitivity and accuracy of non-experts, as well as to define standardized methods for QC.

To fill the gap, an automated, scalable and objective QC environment (QCE) must be conceived for the benefit of the neuro-imaging community. To the best of our knowledge, no supervised Machine Leaning (ML) approach to classifying 3D cortical meshes of neuroimaging pipelines has yet been developed. The aim of this paper is to deploy a novel computational method that can predict "good" and "bad" 3D cortical meshes of the Freesurfer and Civet pipelines (Level-1), as well as to suggest where problem may occur (Level-2), in order to spare neuroscientists' valuable time.

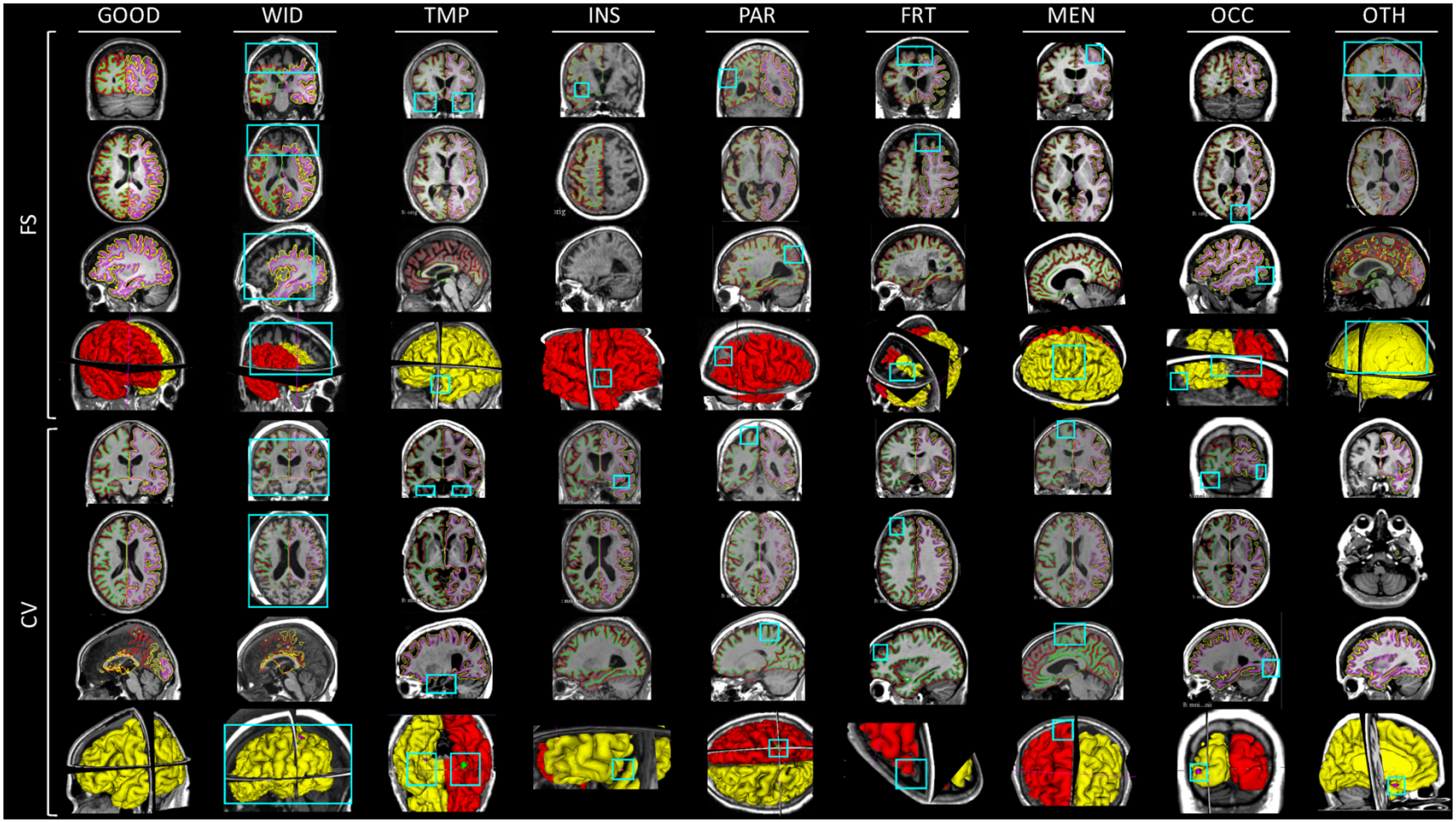


Figure 1 - Snapshots of 3D brain meshes belonging to different QC categories. Green and magenta lines represent white matter surface delineation. Red and yellow lines represent pial surface delineation. The artefacts are surrounded by cyan bounding boxes. In order to not mask problems, the cortical thickness textures are not overimposed on the 3D cortical meshes. FS: Freesurfer; CV: Civet; WID: Widespread problems; TMP: Temporal problems; INS: Insula problems; PAR: Parietal problems; FRT: Frontal problems; MEN: Meninges problem; OCC: Occipital problems; OTH: Other problems.

MATERIALS AND METHODS

Data description

The data used in the preparation of this article are heterogeneous and come from various initiatives, including: ADNI [30], ARWIBO [11], EDSD [27], OASIS [19], and PharmaCOG [13]. High resolution 3D T1-weighted MPRAGE or IRSPGR scans at various magnetic strengths (1.5T or 3.0T) and from different scanner manufacturers (GE, Siemens, Philips), were processed with Freesurfer (v5.3.0) and Civet (v1.1.9) through the neuGRID platform [22, 24]. Overall, 1'582 Freesurfer and 1'692 Civet meshes were used in this study. Figure 2 (Panel A) shows the number of "good" (without artefacts) versus "bad" (with artefacts) meshes for each dataset.

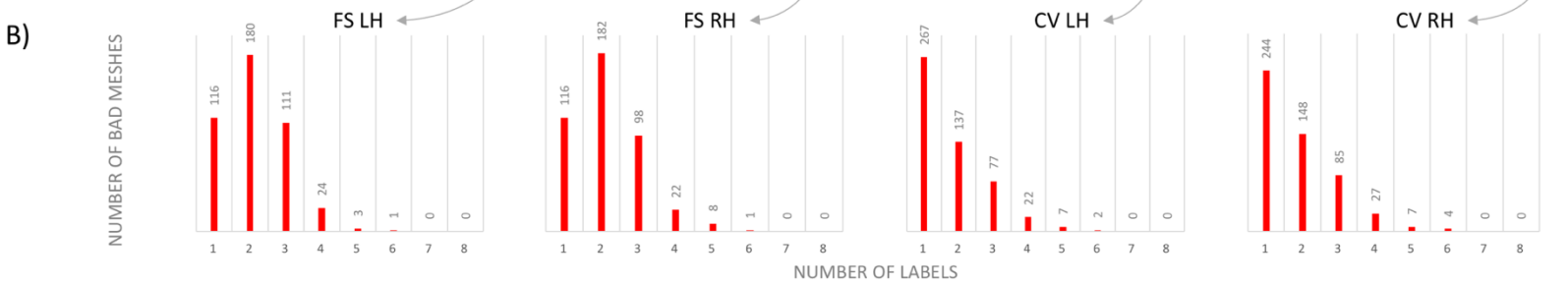
Labelling phase

In order to develop the QCE, a large number of Freesurfer and Civet output was assigned with labels through visual and manual evaluation. Usually, even human raters can have difficulties agreeing on how to label pipeline output. Therefore, in this study, the labels were defined through an evidence-based Delphi panel among three experts that converged on the definition of a "consensus" based on personal experience, the evaluation of 100 common meshes per pipeline, and recursive re-evaluation of choices expressed by other panellists and their justifications thereof. The raters were neuroscientists with more than five years of experience in the neuro-imaging field.

We computed the interrater reliability agreement (IRA) with the Fleiss' kappa coefficient. The IRA values before the Delphi panel for Freesurfer and Civet among the three raters were 0.458 and 0.442, indicating moderate agreement. On the other hand, the IRA values after the Delphi panel for Freesurfer and Civet were 0.87 and 0.93 respectively, indicating almost perfect agreement. Subsequently, all Freesurfer and Civet meshes used were labelled by a single rater. The time needed to accurately label a 3D mesh was, at the least, five minutes per single mesh. Figure 2 reports the benchmark dataset counting meshes with artefacts, labelled as "bad", and allocated in the following eight main categories: (i) widespread problems, (ii) temporal, (iii) insula, (iv) parietal, (v) frontal, (vi) meninges, (vii) occipital and (viii) other problems. The "bad" meshes were coupled with one or more labels (Panel B). The repartition of all the "bad" labels (i.e., 1'834 for Freesurfer; 1'869 for Civet) in the various categories can be found in

A)

	FS				CV			
	LH		RH		LH		RH	
	GOOD	BAD	GOOD	BAD	GOOD	BAD	GOOD	BAD
ADNI	70	63	67	66	38	79	47	70
ARWIBO	52	109	45	116	70	132	70	132
EDSD	41	68	43	66	59	105	64	100
OASIS	151	146	168	129	119	133	107	145
PHARMACOG	42	49	41	50	48	63	43	68
TOT	356	435	364	427	334	512	331	515



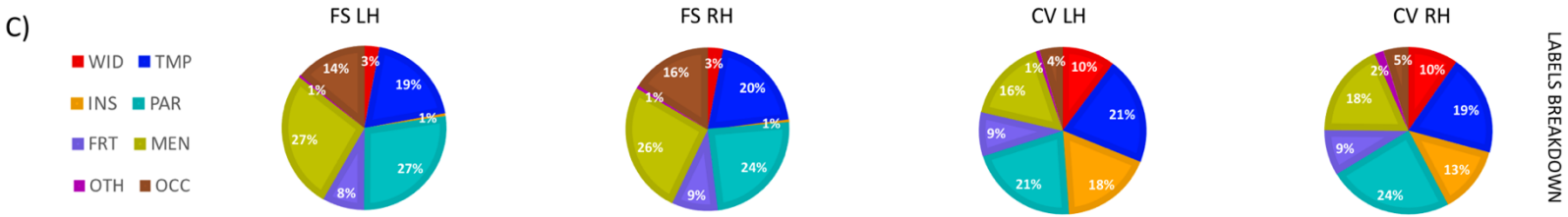


Figure 2 - A) Mesh benchmark dataset composition for Freesurfer and Civet pipelines. The table shows the data source and number of "good" versus "bad" meshes in each hemisphere for each pipeline. B) QC of 3D cortical mesh is a typical multi-label problem where one "bad" 3D mesh can be labelled with more than one artefact/label. C) Labels breakdown per each category in each hemisphere and pipeline. FS: Freesurfer; CV: Civet; WID: Widespread problems; TMP: Temporal problems; INS: Insula problems; PAR: Parietal problems; FRT: Frontal problems; MEN: Meninges problem; OCC: Occipital problems; OTH: Other problems.

Panel C. The benchmark dataset of the 3D meshes generated here, including manual labels, is publicly available (https://neugrid4you.eu/datasets).

Mesh processing and feature extraction

In order to make the manual quality assessment of the rater as well as the automatic features computation easier and effective, we decided to register each mesh in the Talairach reference space. Indeed, 3D meshes of the brain generated by Freesurfer and Civet were in different reference spaces. Affine registration of the Freesurfer meshes was needed, while the Civet meshes were already related to the reference space and no registrations were applied.

Several features based on the derived 3D biomarker descriptors, 2D geometric properties (i.e., image based approach), or 3D distortion measures (i.e., model based approach) [16] were computed (Table 1). A detailed description of the features considered here and their parameter settings will be explained in detail elsewhere.

Main Feature category	**Description**	**Total number of Features extracted**	**Average Features selected in Level-1**		**Average Features selected in Level-2**		**Library / Tool used**
			FS	**CV**	**FS**	**CV**	
Mesh Volume	Computes the volume of white matter mesh, pial mesh, and white/pial mesh ratio.	3	2	2	1	0	Mris_volume
Cortical Thickness Log-Likelihood Ratio Test	Computes the best distribution (from the 79 available in Statsmodels) that fit the cortical thickness, returning the best log-likelihood ratio score.	2	2	1	0	0	Stats models
IRIP-STD in 48 ROIs	Computes the uncertainty of reproducibility error (IRIP) of thickness plus standard deviation (STD) in 48 Harvard Oxford atlas ROIs.	98	49	55	10	10	Numpy/Scipy
Sulcal Shape Descriptors	Computes length, max depth, mean depth, opening, grey matter thickness, average surface of 62 sulci of 3D mesh.	372	207	110	19	11	Aims/Brain Visa (v4.5.0)
3D-Surf-Mask	Computes three overlap measures (i.e., Volume difference, Dice index, Jaccard index) between 3D white matter mesh and 2D WM tissue mask as well as	6	4	2	2	1	FSL/AFNI

	3D pial mesh and 2D GM tissue mask.						
Mesh Structural Distortion Measure	Computes surface to surface distance between two 3D triangle meshes. The measures computed include: (i) 3D Dice coefficient and (ii) 3D intersection/union between white, pial, inflated meshes and *ad-hoc* templates generated from the meshes without artefacts (i.e., "good" meshes).	6	2	1	0	1	3DMetricTool
Zernike Moment	Computes the Zernike moments of 3D mesh surface.	25	12	6	4	4	Mindboggle
Laplace-Beltrami Spectrum	Computes the Laplace-Beltrami spectrum that provides insights into the structure and morphology of shapes.	5	1	2	1	0	Mindboggle
2D Region Property	Computes image properties (area, convex area, eccentricity, Euler number, equivalent diameter, extent, major and minor axis length, orientation, perimeter, solidity, centroid, inertia tensor eigenvalues) of 2D snapshots of the 3D brain meshes on 8 different views (caudal, dorsal, frontal, lateral medial, parietal, rostral, ventral).	184	56	34	8	5	OpenCV-Py /SKimage
3D Unrol Least Square Conformal Map	Computes the 3D mesh parameterization lowering the angle distortion. Once the 3D mesh is unrolled, *2D Region Properties* are computed.	23	8	3	3	2	CGAL
3D Mesh Graph Shortest Path	Computes distances between the centroids of the 48 Harvard Oxford atlas ROIs along the 3D mesh	1127	263	304	39	29	AFNI
2D Graph Property	Computes a simplified 2D graph of flattened mesh and calculates graph properties (density, transitivity, radius, diameter, centrality, closeness, betweenness, estrada, clustering, shortest-paths) of the entire graph as well as of 5 ROIs (frontal, middle temporal, entorhinal, precentral, precuneus areas).	102	7	6	8	6	Networkx
Hash and Hamming Distance	Computes the average Hamming distance of hash values derived from 2D snapshots (caudal, dorsal, frontal, lateral medial,	8	7	5	1	1	Image Hash/photo Hash

	parietal, rostral, ventral) of brain meshes and grayscale cortical thickness versus all "good" meshes.						
CFD	Computes the Navier-Stokes equation, which defines fluid flows (using a steady-state solver for incompressible, turbulent flow) on the 3D brain meshes. These properties include: drag force, lift force, lift to drag ratio, pitching moment and vorticity.	5	4	1	1	0	OpenFOAM
Total		1'966	624	532	97	70	

Table 1 summarizes the main categorical features of 3D meshes computed to train and test the QCE. The average number of features for Freesurfer and Civet are reported in the two different Levels. FS: Freesurfer; CV: Civet; WM: White Matter; GM: grey matter; ROI: region of interest.

Once all the features were computed, the data were centred to null mean and standardized to unit variance in order to prevent the domination of one feature against all others and make the classifier capable of correctly learning from all features as expected. Altogether, we collected 1'966 features per mesh.

Feature selection

To prevent overfitting of the QCE, considering the size of our dataset, we performed feature relevance evaluation and dimensionality reduction. The feature selection was performed within a 10-fold cross-validation method to prevent bias [2]. Inside each one of the cross-validation fold, the feature selection was performed on the training set.

In order to select optimal feature subsets for classification, we used a tree-based feature selection. We considered the mean decrease in Gini index to measure the relevance of each feature [7]. This parameter measures the loss in Gini index on the out-of-bag samples when the feature is removed. The larger the decrease is, the more relevant the feature is. In Level-1, we used 500 tree estimators, while in Level-2 we used 5'000 tree estimators. Features whose importance was greater than the mean Gini loss were kept, the others were discarded. On average, across folds, in Level-1 we selected 624 features for Freesurfer and 532 for Civet, while in Level-2, in average, we considered 97 features for Freesurfer and 70 for Civet. The average numbers of features selected for all folds are reported in Table 1.

QCE structure

The proposed QCE is a two-level classification algorithm (Figure 3). Level-1 plays a central role in the binary classification of "good" and "bad" 3D meshes, while the Level-2 is responsible for constructing the multi-label prediction of bad meshes. We will now introduce the two levels in detail.

Level-1

In Level-1, we used a Random Forest (RF) binary classifier. RF consists of constructing decision trees using randomly selected training samples and features. For classifier modelling, the parameters to be set were: (i) the number of features in the random subset at each tree node, (ii) the number of trees in the forest, (iii) the function used to measure the quality of a split, (iv) the re-weighting strategy to deal with classes imbalance. The first parameter was set to the square root of the total number of features [4]. The other parameters were arranged according to a performance assessment using a nested 10-fold cross-validation grid search strategy. A nested cross-validation consisted in taking one-tenth of the dataset as a validation set, while the rest were used to train and test the classifier in the inner loop with the optimal hyper-parameters. Nested cross-validation was adopted to avoid optimistically biased estimates of performance that result from using the same cross-validation to set the hyper-parameters of the model [5]. The prediction results were kept and the process was repeated for all folds. Our grid search consisted in number of trees equal to {50, 100, 500, 1000, 1'250, 1'500, 1'750, 2'000}; split-function equal to {"Entropy", "Gini"}; re-weighting strategy equal to {"None", "Balanced", "Balanced subsample"}. For both Freesurfer and Civet, a balanced class weight method was chosen. The best final configuration adopted is reported in Table 2.

Level-2

In Level-2 we used a one-vs-the-rest multi-label strategy. Multi-label classification assigns each mesh a set of target labels. This can be thought of as predicting properties of a mesh that are not mutually exclusive. The Level-2 classifier is used to predict multiple labels by fitting a 2D matrix in which cell $[i,j]$ is 1, if mesh i has label j, and 0 otherwise. The classification is performed by an *ad hoc* step of dimensionality reduction, projecting only the first two principal components found by canonical correlation analysis (CCA). The rationale for using CCA is that different artefact typologies of a 3D mesh are often highly correlated to each other (e.g.: "occipital" with "meninges"; "meninges" with "parietal"; "parietal" with "temporal"; etc.). CCA was used to model the association between artefact locations and the 3D mesh properties.

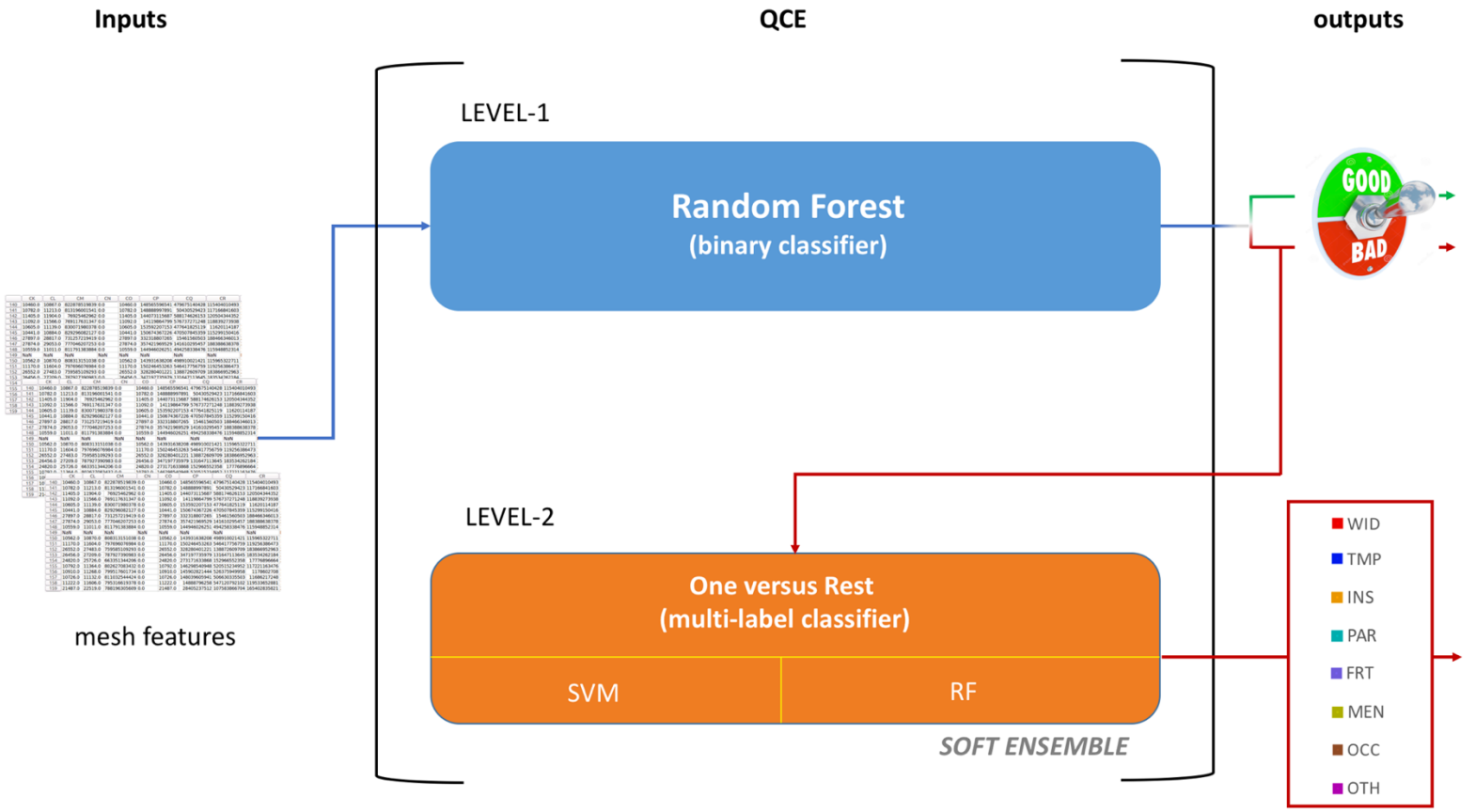


Figure 3 - Simplified overview of the QCE structure. QCE tags the Freesurfer and Civet outputs as "good" or "bad" (Level-1) thanks to a RF estimator. Moreover, QCE detects automatically which portions of the 3D brain meshes are affected by artefacts. It is done exploiting a one-versus-rest approach relying on a soft ensemble multi-label classifier (Level-2). QCE core is designed in Python and scikit-learn [21] as well as on independent external libraries (see Table 1). QCE: Quality Control Environment; SVM: Support Vector Machine; RF: Random Forest. WID: Widespread problems; TMP: Temporal problems; INS: Insula problems; PAR: Parietal problems; FRT: Frontal problems; MEN: Meninges problem; OCC: Occipital problems; OTH: Other problems.

We kept only canonical correlation pairs that were significant. The significance of the first and second canonical correlation pairs of variates were assessed using Wilks' lambda test ($p^{FS}_{Left}<0.001$; $p^{FS}_{Right}<0.001$; $p^{CV}_{Left}<0.05$; $p^{CV}_{Right}<0.001$). Then, an ensemble voting classifier was trained to learn discriminative models for each class. The goal was to combine the prediction of two base estimators in order to improve generalizability and robustness over a single classifier. In this study, a soft-voting combination of Support Vector Machine (SVM) and RF was chosen.

On the one hand, the parameters to be set in the SVM were: (i) the kernel type, (ii) the C value, representing the number of elements accepted in the margin, and (iii) the weights associated with different class labels. On the other hand, the parameters to be set in RF included: (iv) the number of features in the random subset at each tree node, (v) the number of trees in the forest, (vi) the function used to measure the quality of a split; (vii) the weights associated with class labels. As already described in Level-1, the forth parameter was set to the square root of the total number of features. As far as all the other parameters were concerned, these were selected using a nested 10-fold cross-validation grid search strategy. In SVM they were arranged to kernel equal to {"linear", "rbf"}; C equal to {0.01, 0.1, 1, 10, 100}; and re-weighting strategy equal to {"None", "Balanced"}. In RF they were arranged to number of trees equal to {50, 100, 500, 1'000, 1'250, 1'500, 1'750, 2'000}; split-function equal to {"Entropy", "Gini"}; re-weighting strategy equal to {"None", "Balanced", "Balanced subsample"}. Finally, for the soft-voting ensemble, a brute force approach, useful to find the optimal combination of weights, was carried out for both Freesurfer and Civet. The best final configuration adopted is reported in Table 2.

		FS		CV	
		LH	RH	LH	RH
Level-1	Max features	Sqrt(n_features)	Sqrt(n_features)	Sqrt(n_features)	Sqrt(n_features)
	N° estimator	1'000	1'000	1'750	1'750
	Split criterion	Entropy	Entropy	Entropy	Entropy
	Class weight	Balanced subsample	Balanced subsample	Balanced	Balanced
Level-2	Kernel of SVM	RBF	RBF	RBF	RBF
	C of SVM	100	10	100	10
	Class Weight of SVM	None	None	None	None

	Max features of RF	Sqrt(n_features)	Sqrt(n_features)	Sqrt(n_features)	Sqrt(n_features)
	N° estimator of RF	750	250	750	100
	Split criterion of RF	Gini	Entropy	Gini	Gini
	Class weight of RF	Balanced	Balanced subsample	Balanced subsample	Balanced
	Soft Weight [SVM,RF]	[10,8]	[3,2]	[2,5]	[1,17]

Table 2 shows the optimal Level-1 and Level-2 configurations following a nested 10-fold cross validation grid-search strategy. In split criterion, "Gini" and "Entropy" functions seemed to have little effect on the performance of the classifier and were generally consistent respect to each other. In class weight, the "balanced" mode used for Civet adjusted weights inversely proportional to the label frequencies. The "balanced sub-sample" mode used in Freesurfer was similar to the "balanced" method except that weights were computed based on the bootstrap sample for each tree grown. FS: Freesurfer, CV: Civet; LH: left hemisphere; RH: Right hemisphere; Sqrt: square-root; RBF: Radial Basis Function; SVM: Support Vector Machine; RF: Random Forest.

Metrics definition

To evaluate the quality of predictions of our model, we used different metrics detailed in the sections below:

Level-1

$$(1)\quad Accuracy = {}^{1}/_{k} \sum_{i=1}^{k} [(TP_i + TN_i) / (TP_i + FP_i + FN_i + TN_i)]$$

$$(2)\quad Specificity = {}^{1}/_{k} \sum_{i=1}^{k} [TN_i / (TN_i + FP_i)]$$

$$(3)\quad Sensitivity = {}^{1}/_{k} \sum_{i=1}^{k} [TP_i / (TP_i + FN_i)]$$

$$(4)\quad Precision = {}^{1}/_{k} \sum_{i=1}^{k} [TP_i / (TP_i + FP_i)]$$

$$(5)\quad Type\ I\ error = {}^{1}/_{k} \sum_{i=1}^{k} [FP_i / (FP_i + TN_i)]$$

$$(6)\quad Type\ II\ error = {}^{1}/_{k} \sum_{i=1}^{k} [FN_i / (TP_i + FN_i)]$$

Where k is the number of the cross-validation folds, TP is the number of true positive, TN is the number of true negative, FP is the number of false positive, FN is the number of false negative.

Level-2

(1) $Hamming\text{-}Loss(\ y,\ \hat{y}\ ) = \frac{1}{n_{labels}} \sum_{j=0}^{n_{labels}-1} 1(\hat{y}_{\ j} \neq\ y_j)$

(2) $LRAP(y, \hat{f}\ ) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \frac{1}{|y_i|} \sum_{i:y_{ij}=1} \frac{|L_{ij}|}{rank_{ij}}$

with

$L_{ij} = \left\{k: y_{ik} = 1,\ \ \hat{f}_{\ ik} \geq\ \hat{f}_{\ ij}\right\}$;

$rank_{ij} = \left|\left\{k:\ \hat{f}_{\ ik} \geq\ \hat{f}_{\ ij}\right\}\right|$;

$|*|$ is the $\ell 0$-*Norm* or the cardinality of the set;

Where n_{labels} and $n_{samples}$ are the number of labels and samples respectively, y_j is the ground truth, $\hat{y}_{\ j}$ is the QCE prediction for the *j*-th label of a given sample, $\hat{f}$ is the score associated with each label of a sample. The Hamming-Loss is the fraction of labels that are incorrectly predicted [18]. The score obtained is always lower than 1 and the best value is 0.

Label ranking average precision (LRAP) gives the mean fraction of correct positive labels among all positive labels with lower scores for each label [29]. This metric is linked to the average precision function, but is based on the notion of label ranking rather than precision and recall. In the label ranking framework, the objective is to assign a complete preference order of labels for each brain mesh. The obtained score is always strictly greater than 0 and the best value is 1.

RESULTS

QCE evaluation on the whole dataset

Table 3 shows informative metrics to evaluate QCE performances in both pipelines. On the one hand, QCE seemed to be slightly more accurate in discriminating "good" versus "bad" for Freesurfer, although it showed more specificity in Civet. On the other hand, QCE had a better type II error for Freesurfer as well as a better type I error for Civet.

Level-1	ACCURACY	SPECIFICITY	SENSITIVITY	PRECISION	Type I error	Type II error
FS LH	0.89	0.91	0.87	0.89	0.09	0.13
FS RH	0.86	0.87	0.85	0.85	0.13	0.15
CV LH	0.89	0.93	0.84	0.89	0.07	0.16
CV RH	0.83	0.90	0.71	0.83	0.10	0.29

Table 3 - Performance assessment for Level-1. Note that for ACCURACY, SPECIFICITY, PRECISION and SENSITIVITY indices the bigger is the measure-value, the better is the performance. For Type I and Type II errors the smaller is the measure-value, the better is the performance. FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere.

Figure 4 shows the Receiver Operating Characteristic (ROC) curves used to discriminate "good" and "bad" meshes. The corresponding areas under the curves (AUC) were computed and statistically compared setting the threshold for significance at a p value of 0.05. In both pipelines, QCE showed statistical difference ($p<0.05$) between the AUCs derived with meshes of the left and those derived with the right hemisphere. In Freesurfer, the discriminative analysis yielded 0.96 in the left and 0.94 in the right hemisphere ($z = +2.09$); while in Civet it yielded 0.96 and 0.91 respectively ($z = +4.17$). QCE performed better in terms of classification accuracy on left meshes although it is characterised by high AUCs in both Freesurfer's and Civet's right meshes. QCE looked similarly robust for both pipelines in Level-1.

In machine learning (ML), performance evaluation of multi-label classification differs from that of classical binary classification as each sample may have more labels simultaneously [25].

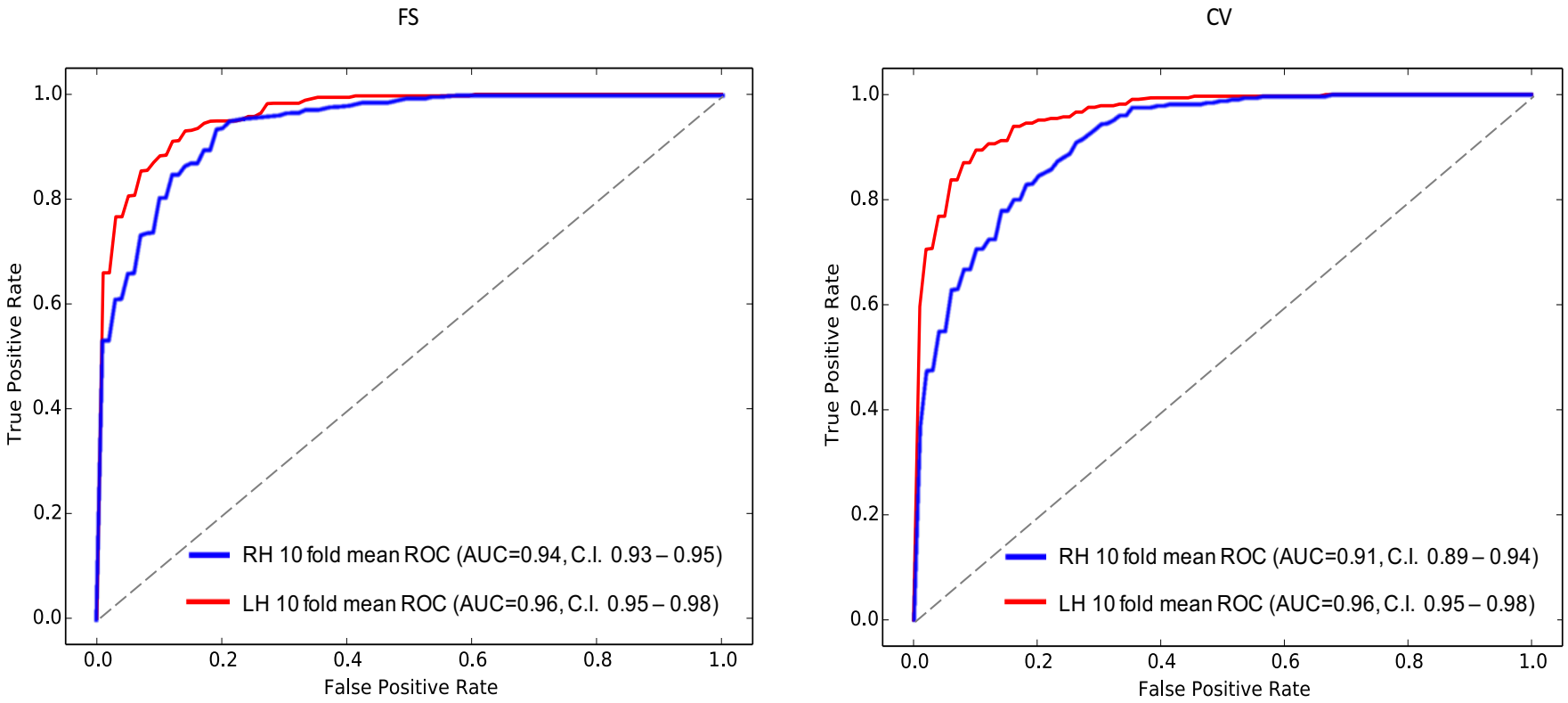


Figure 4 - Receiving Operator Characteristic (ROC) curves showing the performances of the QCE in classifying "good" versus "bad" 3D meshes of Freesurfer (left) and Civet (right). FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere; AUC: Area Under the Curve; C.I.: 95% Confidence Interval.

Table 4 shows two common measures used to evaluate the QCE performance in Level-2. QCE shows a good Hamming-Loss (HL) in labelling both left and right 3D bad mesh subtypes in Freesurfer as well as in Civet.

As far as the LRAP index is concerned, this is a strict metric not characterized by over prediction or under prediction. It is therefore difficult to obtain very high overall LRAP considering eight multiple labels. QCE showed a slightly higher LRAP index in Freesurfer than Civet in both the left and right hemispheres.

Level-2	HL	LRAP
FS LH	0.19	0.62
FS RH	0.21	0.59
CV LH	0.22	0.51
CV RH	0.23	0.51

Table 4 - Performance assessment for Level-2. FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere; HL: Hamming-Loss; LRAP: Label ranking average precision.

For the benefit of the reader, Figure 5 graphically shows the QCE in classifying 3D meshes as multiple combinations of atomic labels: (a) widespread problems, (b) temporal, (c) insula, (d) parietal, (e) frontal, (f) meninges, (g) occipital and (h) other problems.

QCE evaluation on independent datasets

To evaluate the prediction performance of our classifier further, we applied a leave one-group-out strategy. Essentially, we were interested to know whether the QCE trained on a particular set of data-groups, generalizes well to the unseen group. To measure this, one dataset at a time was held out as the testing dataset. The remaining datasets were used to train our QCE each time from scratch. Following this procedure, we then tested the prediction performance on the testing dataset. In each independent dataset, the performance details obtained by QCE can be found in Table 5 and Table 6.

As far as the Level-1 is concerned, the QCE had generally lower sensitivity than that reached for the whole dataset. In particular, the loss was more pronounced in Freesurfer than Civet especially in the ADNI and OASIS datasets. The type II errors doubled in EDSD and Pharma-

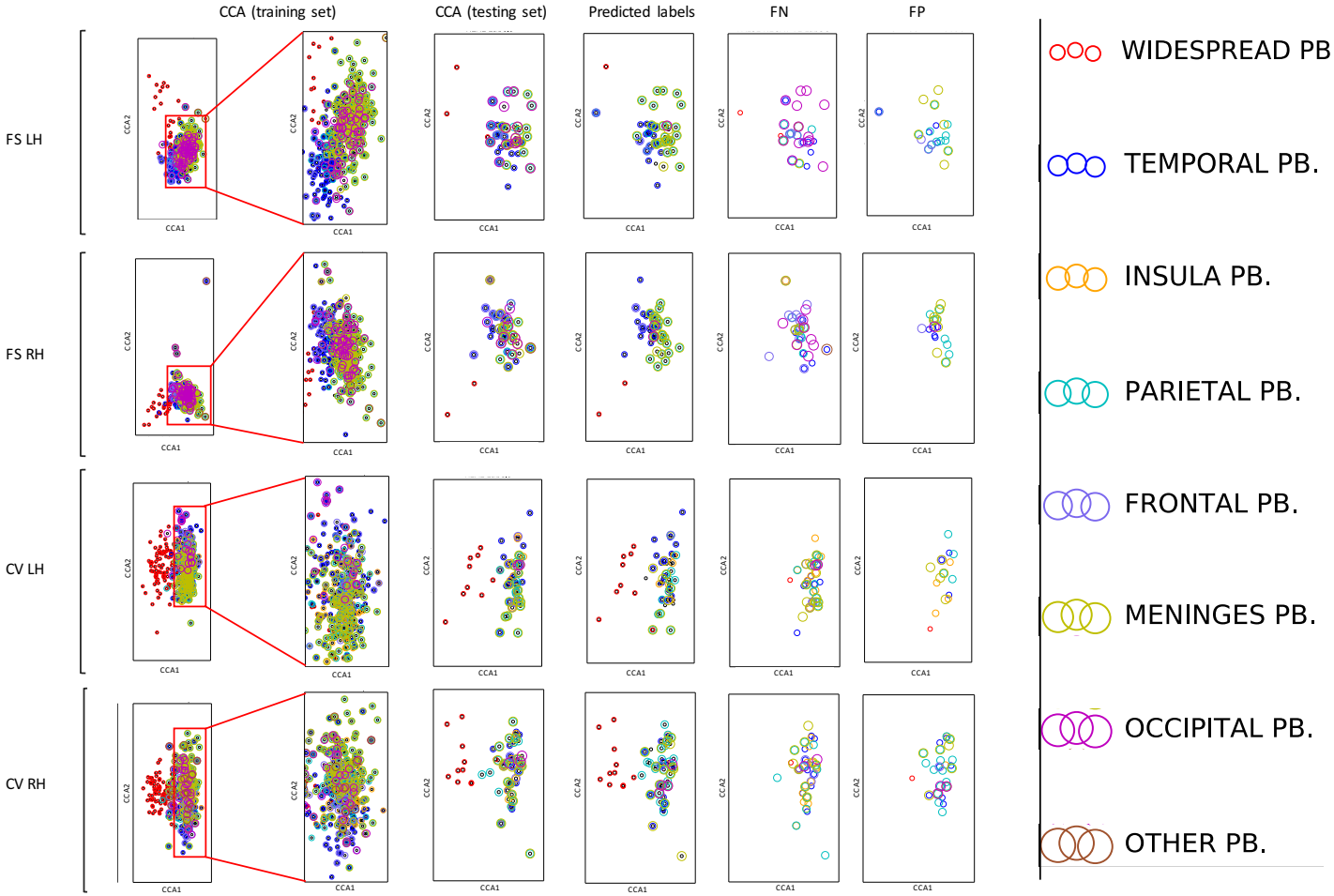


Figure 5 represents one k-fold of the QCE 10-fold cross-validation of the Level-2 for both Civet and Freesurfer. The 3D meshes that are affected by multiple artefacts are plotted surrounded by concentric colored circles. The predicted labels are quite consistent with the real labels tagged by expert. The QCE incorrect labels are shown in the last two columns. CCA is used to perform a supervised dimensionality reduction. A complete overview of the 10-fold cross validation test-sets is provided in supplementary material (Figure S1). FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere. CCA: Canonical Correlation Analysis. FN: False negative Labels (labels not found, that are correct). FP: False Positive labels (labels found, that are incorrect). PB: problems.

Figure S1 represents each test-set of the QCE assessed in the 10-fold cross-validation for Freesurfer and Civet pipelines. Real (i.e.: ground truth), predicted, false negative, and false positive labels are shown. FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere.

COG and tripled in ADNI and OASIS for Freesurfer, while they doubled in EDSD and OASIS for Civet. The accuracy of the QCE on both hemispheres was overall comparable to the relative performance of the QCE assessed for the whole dataset. The AUC of the QCE in both pipeline, as well as in each hemisphere, was always greater than 0.86, reaching values of 0.99 in some independent dataset. This indicates QCE in Level-1 was not over-estimated.

As far as the Level-2 was concerned, the overall performance in terms of HL and LRAP on each independent dataset were strictly in line with the scores reached by the QCE in the whole cross-validated dataset. This indicated that QCE in Level-2 is well-balanced and not over-fitted.

Level-1	Independent dataset test	ACCURACY	SPECIFICITY	SENSITIVITY	PRECISION	Type I error	Type II error	AUC
FS LH	ADNI	0.86	0.99	0.73	0.99	0.01	0.27	0.98
FS RH		0.73	0.96	0.49	0.92	0.04	0.51	0.90
CV LH		0.99	0.98	0.99	0.97	0.01	0.01	0.99
CV RH		0.97	0.99	0.93	0.99	0.01	0.06	0.99
FS LH	ARWIBO	0.84	0.83	0.85	0.71	0.17	0.15	0.92
FS RH		0.77	0.77	0.78	0.56	0.23	0.22	0.86
CV LH		0.91	0.93	0.87	0.87	0.07	0.13	0.96
CV RH		0.78	0.89	0.59	0.72	0.12	0.41	0.86
FS LH	EDSD	0.90	0.99	0.73	0.99	0.01	0.27	0.96
FS RH		0.83	0.91	0.72	0.84	0.09	0.28	0.89
CV LH		0.85	0.95	0.68	0.89	0.05	0.32	0.94
CV RH		0.75	0.95	0.44	0.85	0.05	0.56	0.89
FS LH	OASIS	0.82	0.95	0.71	0.93	0.05	0.29	0.94
FS RH		0.71	0.91	0.57	0.89	0.09	0.43	0.89
CV LH		0.83	0.97	0.68	0.95	0.03	0.32	0.95
CV RH		0.73	0.92	0.48	0.81	0.08	0.52	0.85
FS LH	PHARMACOG	0.86	0.98	0.71	0.97	0.02	0.29	0.98
FS RH		0.86	0.94	0.76	0.91	0.06	0.24	0.93
CV LH		0.93	0.98	0.85	0.98	0.02	0.15	0.98
CV RH		0.78	0.74	0.86	0.67	0.26	0.14	0.87

Table 5- Performance of the QCE for Level-1 on independent datasets. Each training set is constituted by all the samples except the ones related to the independent group (i.e.: dataset test). FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere; AUC: Area Under the Curve.

Level-2	Independent dataset test	HL	LRAP
FS LH	ADNI	0.22	0.59
FS RH		0.19	0.60
CV LH		0.20	0.51
CV RH		0.24	0.46
FS LH	ARWIBO	0.20	0.62
FS RH		0.16	0.67
CV LH		0.26	0.49
CV RH		0.29	0.49
FS LH	EDSD	0.21	0.56
FS RH		0.23	0.56
CV LH		0.24	0.50
CV RH		0.25	0.50
FS LH	OASIS	0.20	0.60
FS RH		0.24	0.53
CV LH		0.18	0.54
CV RH		0.21	0.52
FS LH	PHARMACOG	0.20	0.62
FS RH		0.20	0.64
CV LH		0.24	0.53
CV RH		0.20	0.56

Table 6 - Performance of the QCE for Level-2 on independent datasets. Each training set is constituted by all the samples except the ones related to the independent group (i.e.: dataset test). FS: Freesurfer; CV: Civet; RH: Right hemisphere; LH: Left hemisphere; HL: Hamming-Loss; LRAP: Label ranking average precision.

DISCUSSION

The aim of this paper was to assess whether an automatic ML tool could perform QC of MR-based cortical surface derived by two well-known pipelines, i.e., Freesurfer and Civet. More specifically, our purpose was to investigate whether error detection and its related localization can be reliably handled by a supervised classifier. This is not an easy task because it requires: (i) high sensitivity and (ii) sufficiently low type II error.

According to our experiments, in Level-1 the QCE produced excellent results with high precision and acceptable sensitivity in line with those achieved by trained experts for both pipelines.

The current version of the QCE relies on various 3D mesh features, mainly based on texture, statistics and geometric context information. The best results for Level-1 binary classification were primarily obtained with IRIP-STD, HamHash, Sulci descriptors and the 3D shortest paths in terms of relevance and number of features selected.

In Level-2, Zernike momentum and Beltrami spectrum features were the most important. On the contrary, we should note that LogLH features have never been selected by either the Freesurfer or Civet pipelines, indicating that these properties were of no relevance in the multi-label tagging task.

The present work is characterized by large and variegated multi-diagnosis, multi-cohort and multi-site datasets. This is the best situation to test QCE performance even on output derived from poor quality input images, from different manufacturer's scanners and from different field strengths. As such, AUCs between 0.91 and 0.96 in Level-1 were especially good. However, QCE produced AUCs statistically lower in the discriminant analysis of the right hemispheric meshes versus the left ones. This was in line with the IRA of the panellists (IRA^{FS}_{Right}= +0.824; IRA^{FS}_{Left}= +0.912; IRA^{CV}_{Right}= +0.916; IRA^{CV}_{Left}= +0.940). A probable explanation could be the different and suboptimal framework of artefacts visualisation adopted (i.e.: different colours for display gray/white matters of the left/right 3D meshes) that could simplify or get more difficult the artefacts detection at naked eye. This might influence the final label-set as well. Thus, in future study, this aspect should be further investigated.

In Level-2, the QCE performed sufficiently well during the multi-label validation phase, although it did have some missing or incorrect labels. For Freesurfer, QCE tended to miss occipital and frontal problems, as well as to misclassify temporal and parietal artefacts. In Civet, the situation was similar but insular labels were also often incorrectly identified. This was probably due to fuzzy clusters obtained after the supervised CCA dimensionality

reduction. In future studies, we aim to improve the QCE type I and type II errors exploring new ML algorithms, feature dimensionality reduction techniques as well as adding new informative 3D mesh spatial information for both Level-1 and Level-2.

Interestingly, in Level-1, the QCE configuration and its hyper-parameters were much more homogeneous between the left and right hemispheric meshes than in Level-2. This occurred for both pipelines. On the other hand, in the training set of Level-2, the clouds of different mesh problems were not clearly distinguishable and well-segregated, showing different conformations for the two hemispheres. For these reasons, the Level-2 hyper-parameters configuration diverged considerably from one hemisphere to the other. However, this is in line with the expected capabilities of the system to fit, as much as possible, the clouds of each composite label. Hence, QCE is good at adjusting to different situations providing overall comparable and robust results for both Freesurfer and Civet. Furthermore, the differences between the two pipelines in the HL and LRAP metrics can be partially explained in term of: (i) different representativeness of the bad artefacts, more uniformly distributed in Civet than in Freesurfer; (ii) different numbers of artefacts per mesh, scarcer multiple labels in Civet than in Freesurfer.

Notably, QCE still performed robustly on the five independent datasets assessed and more importantly, it was not over-fitted. In Level-1, the standard deviations of QCE accuracies were small: 0.03 and 0.06 in the left and right hemisphere respectively for Freesurfer; 0.06 and 0.10 in the left and right hemisphere respectively for Civet. In Level-2, from the five independent datasets, the standard deviation of HL and LRAP never exceeded 0.06 for Freesurfer and 0.04 for Civet.

QCE is an ideal solution in terms of scalability. It can classify thousands of 3D cortical meshes quicker and with higher efficiency than an expert, that is unavoidably prone to subjective errors during the visual inspection, can do.

In this first study, we used annotations based on a consensus of just three experts as a reference to measure QCE performance. Although it is a common approach for validation in the medical image processing community, future work should include reference data from many more observers. Proceeding along these lines, the definition of an international harmonized protocol for 3D cortical mesh problems detection, as well as publicly well-annotated database, could serve as a reference standard for the advancement of future automated QCE systems.

Certain restrictions should be considered in the interpretation of these results. First, the QCE needs to be further tested with other recent cortical delineation techniques, such as: DiReCT

[8], CLADA [20], and MILXCTE [1] to see if comparable results can be obtained with meshes produced by different algorithms. Second, although in a multi-label dataset, the frequency of labels is often not even, we could deal with the multi-label imbalance with some very recent approaches such as: LP-RUS or LP-ROS [6]. Third, even if the main features of the QCE can be parallelized in order to reduce the time needed to collect numerical features from the cortical mesh, some (e.g., 3D-surf-mask, Zernike momentum, Laplace-Beltrami spectrum, 2D graph properties, and Computational Fluid Dynamics) are very demanding and need few hours to complete. Thus, to overcome this limitation, grid/cloud e-infrastructures are mandatory to make the 3D mesh processing, features extraction and QCE execution slighter and faster. In order to maximize software distribution, a more streamlined QCE version could be compiled with less dependences to compute just the most informative features, as well as enabling neuroscientists to run QCE locally on their own resources.

CONCLUSION

The detection of 3D cortical mesh errors in an objective and automatic manner is now possible. Thanks to the combination of innovative ML approaches, we have developed and endorsed a system capable of predicting whether a 3D cortical mesh is correct or affected by artefacts and in the latter case, to suggest where problems may be.

In this article, a new benchmark QC dataset has been constructed that contains 3D cortical meshes from Freesurfer and Civet with both single and multiple labels.

Future work will focus on increasing the performance of our method in artefact detection, so that QC tasks can be handled entirely and exclusively using an automated approach.

Given its scalability and efficacy, QCE represents an opportunity to process data more easily and quickly, allowing neuroscientists to spend their valuable time completely on data analysis instead of spending resources in tedious QC tasks.

REFERENCES

1. Acosta O, Fripp J, Doré V, Bourgeat P, Favreau JM, et al. Cortical surface mapping using topology correction, partial flattening and 3D shape context-based non-rigid registration for use in quantifying atrophy in Alzheimer's disease. J Neurosci Methods. 2012 205(1): 96–109. doi: 10.1016/j.jneumeth.2011.12.011
2. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A. 2002 May 14;99(10):6562-6.
3. Amunts K, Ebell C, Muller J, Telefont M, Knoll A, Lippert T. The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain. Neuron. 2016 Nov 2;92(3):574-581. doi: 10.1016/j.neuron.2016.10.046.
4. Breiman L. Bagging Predictors. Machine Learning. 1996 24: pp.123–140. doi:10.1023/A:1018054314350
5. Cawley GC, Talbot NLC. Over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research. 2010 11:2079-2107.
6. Charte F, Rivera AJ, del Jesus MJ, Herrera F. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing. 2014 pp.150-160. doi: 10.1007/978-3-642-40846-5{_}16
7. Cutler A, Cutler DR, Stevens JR. Random Forests. Ensemble Machine Learning: Methods and Applications. Springer US. 2012. pp: 157–175. doi: 10.1007/978-1-4419-9326-7_5
8. Das SR, Avants BB, Grossman M, Gee JC. Registration based cortical thickness measurement. Neuroimage. 2009 45: 867–879. doi: 10.1016/j.neuroimage.2008.12.016
9. Ducharme S, Albaugh MD, Nguyen TV, Hudziak JJ, Mateos-Pérez JM, Labbe A, Evans AC, Karama S; Brain Development Cooperative Group. Trajectories of cortical thickness maturation in normal brain development--The importance of quality control procedures. Neuroimage. 2016 Jan 15;125:267-79. doi: 10.1016/j.neuroimage.2015.10.010. Epub 2015 Oct 14.
10. Fischl B, Dale AM. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci. 2000 USA 97(20): 11050–11055.
11. Frisoni GB, Prestia A, Zanetti O, Galluzzi S, Romano M, Cotelli M, Gennarelli M, Binetti G, Bocchio L, Paghera B, Amicucci G, Bonetti M, Benussi L, Ghidoni R, Geroldi C. Markers of Alzheimer's disease in a population attending a memory clinic. Alzheimers Dement. 2009 Jul;5(4):307-17.
12. Frisoni GB, Redolfi A, Manset D, Rousseau MÉ, Toga A, Evans AC. Virtual imaging laboratories for marker discovery in neurodegenerative diseases. Nat Rev Neurol. 2011 Jul 5;7(8):429-38. doi: 10.1038/nrneurol.2011.99. Review.
13. Galluzzi S, Marizzoni M, Babiloni C, Albani D, Antelmi L, Bagnoli C, Bartres-Faz D, Cordone S, Didic M, Farotti L, Fiedler U, Forloni G, Girtler N, Hensch T, Jovicich J, Leeuwis A, Marra C, Molinuevo JL, Nobili F, Pariente J, Parnetti L, Payoux P, Del Percio C, Ranjeva JP, Rolandi E, Rossini PM, Schönknecht P, Soricelli A, Tsolaki M, Visser PJ, Wiltfang J, Richardson JC, Bordet R, Blin O, Frisoni GB; PharmaCog Consortium. Clinical and biomarker profiling of prodromal Alzheimer's disease in workpackage 5 of the Innovative Medicines Initiative PharmaCog project: a 'European ADNI study'. J Intern Med. 2016 Jun; 279(6):576-91.
14. Ganz M, Kondermann D, Andrulis J, Knudsen GM, Maier-Hein L. Crowdsourcing for error detection in

cortical surface delineations. Int J Comput Assist Radiol Surg. 2017 Jan;12(1):161-166. doi: 10.1007/s11548-016-1445-9. Epub 2016 Jun 27.

15. Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab'bagh Y, MacDonald D, Lee JM, Kim SI, Evans AC. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. Neuroimage. 2005 Aug 1;27(1):210-21.
16. Lavoué G, Liu H, Myszkowski K, Lin W. Quality Assessment and Perception in Computer Graphics. IEEE Computer Graphics and Applications, vol. 36, no. 4, pp. 21-22, July-Aug. 2016. doi: 10.1109/MCG.2016.72
17. Lee JK, Lee JM, Kim JS, Kim IY, Evans AC, Kim SI. A novel quantitative cross-validation of different cortical surface reconstruction algorithms using MRI phantom. Neuroimage. 2006 Jun;31(2):572-84. Epub 2006 Feb 24.
18. Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. Pattern Recognition. 2012 (45)9: 3084-3104. doi: http://doi.org/10.1016/j.patcog.2012.03.004.
19. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. Journal of Cognitive Neuroscience. 2010 Dec. 22 (12): 2677 - 2684.
20. Nakamura K, Fox R, Fisher E. CLADA: Cortical longitudinal atrophy detection algorithm. Neuroimage. 2011 54(1): 278–289. doi: 10.1016/j.neuroimage.2010.07.052
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot, Duchesnay É. Scikit-learn: Machine Learning in Python. 2011 JMLR 12, pp. 2825-2830.
22. Redolfi A, McClatchey R, Anjum A, Zijdenbos A, Manset D, Barkhof F, Spenger C, Legré Y, Wahlund LO, Barattieri di San Pietro C, Frisoni GB. Grid infrastructures for computational neuroscience: the neuGRID example. Future Neurology. 2009, Vol. 4, No. 6, Pages 703-722, doi:10.2217/fnl.09.53.
23. Redolfi A, Bosco P, Manset D, Frisoni GB; neuGRID consortium. Brain investigation and brain conceptualization. Funct Neurol. 2013 Jul-Sep;28(3):175-90. doi: 10.11138/FNeur/2013.28.3.175. Review.
24. Redolfi A, Manset D, Barkhof F, Wahlund LO, Glatard T, Mangin JF, Frisoni GB; neuGRID Consortium, for the Alzheimer's Disease Neuroimaging Initiative. Head-to-head comparison of two popular cortical thickness extraction algorithms: a cross-sectional and longitudinal study. PLoS One. 2015 Mar 17;10(3):e0117692. doi: 10.1371/journal.pone.0117692.
25. Sechidis K, Tsoumakas G, and Vlahavas I. On the stratification of multilabel data. In Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III, ECML PKDD'11, pp. 145–158, Berlin, Heidelberg, 2011. Springer-Verlag.
26. Sherif T, Rioux P, Rousseau ME, Kassis N, Beck N, Adalat R, Das S, Glatard T, Evans AC. CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. Front Neuroinform. 2014 May 21;8:54. doi: 10.3389/fninf.2014.00054. eCollection 2014.
27. Teipel SJ, Wegrzyn M, Meindl T, Frisoni G, Bokde AL, Fellgiebel A, Filippi M, Hampel H, Klöppel S, Hauenstein K, Ewers M. Anatomical MRI and DTI in the diagnosis of Alzheimer's disease: a European

multicenter study. Journal of Alzheimer's Disease. 2012 Jan 1;31(s3).

28. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, Schumann G, Franke B, Wright MJ, Martin NG, Agartz I, Alda M, Alhusaini S, Almasy L, Almeida J, Alpert K, Andreasen NC, Andreassen OA, Apostolova LG, Appel K, Armstrong NJ, Aribisala B, Bastin ME, Bauer M, Bearden CE, Bergmann O, Binder EB, Blangero J, Bockholt HJ, Bøen E, Bois C, Boomsma DI, Booth T, Bowman IJ, Bralten J, Brouwer RM, Brunner HG, Brohawn DG, Buckner RL, Buitelaar J, Bulayeva K, Bustillo JR, Calhoun VD, Cannon DM, Cantor RM, Carless MA, Caseras X, Cavalleri GL, Chakravarty MM, Chang KD, Ching CR, Christoforou A, Cichon S, Clark VP, Conrod P, Coppola G, Crespo-Facorro B, Curran JE, Czisch M, Deary IJ, de Geus EJ, den Braber A, Delvecchio G, Depondt C, de Haan L, de Zubicaray GI, Dima D, Dimitrova R, Djurovic S, Dong H, Donohoe G, Duggirala R, Dyer TD, Ehrlich S, Ekman CJ, Elvsåshagen T, Emsell L, Erk S, Espeseth T, Fagerness J, Fears S, Fedko I, Fernández G, Fisher SE, Foroud T, Fox PT, Francks C, Frangou S, Frey EM, Frodl T, Frouin V, Garavan H, Giddaluru S, Glahn DC, Godlewska B, Goldstein RZ, Gollub RL, Grabe HJ, Grimm O, Gruber O, Guadalupe T, Gur RE, Gur RC, Göring HH, Hagenaars S, Hajek T, Hall GB, Hall J, Hardy J, Hartman CA, Hass J, Hatton SN, Haukvik UK, Hegenscheid K, Heinz A, Hickie IB, Ho BC, Hoehn D, Hoekstra PJ, Hollinshead M, Holmes AJ, Homuth G, Hoogman M, Hong LE, Hosten N, Hottenga JJ, Hulshoff Pol HE, Hwang KS, Jack CR Jr, Jenkinson M, Johnston C, Jönsson EG, Kahn RS, Kasperaviciute D, Kelly S, Kim S, Kochunov P, Koenders L, Krämer B, Kwok JB, Lagopoulos J, Laje G, Landen M, Landman BA, Lauriello J, Lawrie SM, Lee PH, Le Hellard S, Lemaître H, Leonardo CD, Li CS, Liberg B, Liewald DC, Liu X, Lopez LM, Loth E, Lourdusamy A, Luciano M, Macciardi F, Machielsen MW, Macqueen GM, Malt UF, Mandl R, Manoach DS, Martinot JL, Matarin M, Mather KA, Mattheisen M, Mattingsdal M, Meyer-Lindenberg A, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meisenzahl E, Melle I, Milaneschi Y, Mohnke S, Montgomery GW, Morris DW, Moses EK, Mueller BA, Muñoz Maniega S, Mühleisen TW, Müller-Myhsok B, Mwangi B, Nauck M, Nho K, Nichols TE, Nilsson LG, Nugent AC, Nyberg L, Olvera RL, Oosterlaan J, Ophoff RA, Pandolfo M, Papalampropoulou-Tsiridou M, Papmeyer M, Paus T, Pausova Z, Pearlson GD, Penninx BW, Peterson CP, Pfennig A, Phillips M, Pike GB, Poline JB, Potkin SG, Pütz B, Ramasamy A, Rasmussen J, Rietschel M, Rijpkema M, Risacher SL, Roffman JL, Roiz-Santiañez R, Romanczuk-Seiferth N, Rose EJ, Royle NA, Rujescu D, Ryten M, Sachdev PS, Salami A, Satterthwaite TD, Savitz J, Saykin AJ, Scanlon C, Schmaal L, Schnack HG, Schork AJ, Schulz SC, Schür R, Seidman L, Shen L, Shoemaker JM, Simmons A, Sisodiya SM, Smith C, Smoller JW, Soares JC, Sponheim SR, Sprooten E, Starr JM, Steen VM, Strakowski S, Strike L, Sussmann J, Sämann PG, Teumer A, Toga AW, Tordesillas-Gutierrez D, Trabzuni D, Trost S, Turner J, Van den Heuvel M, van der Wee NJ, van Eijk K, van Erp TG, van Haren NE, van 't Ent D, van Tol MJ, Valdés Hernández MC, Veltman DJ, Versace A, Völzke H, Walker R, Walter H, Wang L, Wardlaw JM, Weale ME, Weiner MW, Wen W, Westlye LT, Whalley HC, Whelan CD, White T, Winkler AM, Wittfeld K, Woldehawariat G, Wolf C, Zilles D, Zwiers MP, Thalamuthu A, Schofield PR, Freimer NB, Lawrence NS, Drevets W; Alzheimer's Disease Neuroimaging Initiative, EPIGEN Consortium, IMAGEN Consortium, Saguenay Youth Study (SYS) Group. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 2014; 8(2): 153–182. Published online 2014 Jan 8. doi: 10.1007/s11682-013-9269-5.

29. Wei C, Liu T, Lan Y, Ma Z, Hang L; Ranking Measures and Loss Functions in Learning to Rank. Advances in Neural Information Processing Systems 22 - Curran Associates, Inc. 2009. NIPS2009_3708: 315—323.
30. Weiner MW, Aisen PS, Jack CR, Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA, Toga A. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimer's & Dementia. 2010 May 31;6(3):202-11.

Chapter 5

Conclusion

GENERAL DISCUSSION

Neuroimaging research, by its very nature, is data intensive [1]. In this thesis, we proved that neuroimaging is an example of discovery-oriented science, wherein it is possible to systematically extract and study patterns of brain structure across thousands of subjects and dozens of studies, resulting in new knowledge.

The computational environments to support this advanced form of brain research are still maturing. Despite their common vision, e-infrastructures today available have been designed and developed at different time points and in different scientific contexts, to address specific contingent needs. As a consequence, while sharing many commonalities, they also present significant differences. Some of them are more close-fitted to a data-federation approach (such as GAAIN and HBP-MIP) adopting the slogan "it is easier to move algorithms than data across centres" [2]; others (such as CATI, CBRAIN, LONI, and neuGRID) are testimonials of a data-sharing schema through a more centralised approach.

All these e-infrastructures offer computing power and storage capacity that benefit from the combination of different resources, such as grids, regular HPCs, and public/private clouds, to increase their overall performance.

The general setting is optimal to start thinking about new multimodal and, possibly, multiscale brain models such as those developed in the context of the "Human Brain Project" or the "Brain Activity Map" initiatives in Europe and USA, respectively (Chapter 2).

Waiting for this to happen, it is nonetheless possible to exploit the large amount of data already stored in e-infrastructures to test new hypotheses and better characterize the AD pathology. One of the study we were interested in was the relative efficacy of the algorithms used to quantify the brain's cortical thickness. In this sense, we performed a head-to-head comparison of Freesurfer and Civet (two very popular pipelines) to validate one of the most promising AD biomarker (Chapter 3).

Our purpose was to understand which of the two tools was the most accurate for tracking disease evolution. Despite being a computationally expensive task, we succeeded in obtaining results in less than 10 weeks running the jobs on the neuGRID and VIP resources (to be compared with the 5-year timeframe we would have needed on a single mono-core computer). In particular, this study represents a first attempt to verify the mutual strengths and weaknesses of Civet and Freesurfer at a precision level of the single voxel. This was possible thanks to an

innovative hybrid template or, more specifically, to the exploitation of a 3D Gradient Vector Flow (GVF) and Closest Point Search (CPS) techniques between the two standard pipelines' templates, which allowed for the definition of a mutual vertex by vertex correspondence between Freesurfer's and Civet's meshes.

Cross-sectionally, both Civet's and Freesurfer's performances changed according to the disease stage. Both tools detected the characteristic signature of AD (i.e.: the posterior cingulate, temporal lobe and frontal gyrus, medial temporal, retrosplenial, and lateral temporal regions).

Longitudinally, the main results of the study were: (i) in pMCI Civet was able to highlight a characteristic atrophic pattern involving expected temporal areas, such as the inferior margin of central gyrus and extended lateral frontal-parietal areas, as expected; (ii) in pMCI Freesurfer was not able to find as many regions as those detected by Civet with the same significance and extension (i.e.: orbital, triangular, and opercular portion of the inferior frontal gyrus, transverse-temporal and mesial part of the superior frontal cortex, inferior parietal cortex, the superior temporal gyrus); (iii) the Civet's more representative cortical signature suggests that this tool can detect the typical atrophic patterns in subject that will convert to AD within 2 years more efficiently than Freesurfer.

Additionally, such systematic and analytical research-study, was propaedeutic for the implementation of an automatic QCE (Chapter 4). In other words, the study set the bases to develop a ML tools capable to automatically classify 3D cortical meshes in different categories using specific cortical properties.

For both pipelines, our QCE showed brilliant results with high precision and acceptable sensitivity in classifying 3D meshes as "good" or "bad", in line with the performance of trained experts. The AUCs for the binary classification yielded the following results: AUC^{FS}_{Left}= 0.96; AUC^{FS}_{Right}=0.94; AUC^{CV}_{Left}=0.96; AUC^{CV}_{Right}=0.91. If a 3D mesh is classified as "bad", the QCE suggests where artefacts are located according to the following classification schema: (i) widespread problems, (ii) temporal, (iii) insula, (iv) parietal, (v) frontal, (vi) meninges, (vii) occipital, and (viii) other problems. This is a typical multi-label classification problem addressed with a combination of machine learning tools, i.e.: soft-ensemble classifier based on Support Vector Machine (SVM) and Random Forest (RF). Our QCE relies on different mesh features based on texture, thickness statistics, and geometric context information.

QCE represents a milestone for every neuroscientist in the field of neuroimaging. For the first time, a fully automated and objective method for the quality assurance is available to be plugged at the end of any MRI cortical thickness analysis.
The most immediate and tangible results are: (a) the considerable saving of time; (b) the reduction of subjective errors; (c) the higher efficiency and scalability in term of QC assessments and sanity check per unit-time (e.g., hour, day, week, month, year), this latter being absolutely mandatory in the current big-data era. QCE represents the final product of this thesis.

So, the "red thread" for neuroscientists (i.e., e-infrastructure, cortical mesh segmentation, quality control environment) has been finally tracked.

Methodological considerations

Strengths

One of the main strengths of our study was the use of neuGRID, a dedicated e-infrastructure. This was possible thanks to a grant won in the context of the European call DG-CONNECT (agreement no. 283562 FP7/2007-2013).
In addition to high computational resources, we used a large dataset to compare and assess the cortical thickness algorithms. We included participants across the whole cognitive spectrum in both cross-sectional and longitudinal series.
Further, a multi-cohort dataset was assembled for the development of the QCE on huge and heterogeneous amount of T13D MRI scans coming from different countries in America and Europe which increased the generalizability of our results.

Limitations

Our studies presents nonetheless several limitations. The thought that, from now on, data will simply be processed "in the cloud" is somewhat naïve. The grid/cloud e-infrastructures physically exist, somewhere, and system failures can always happen. Moreover, physical resources are not infinite, by definition, and are costly, too.
As far as the "3D GVF – CPS" methodology is concerned, further tests on the newly released version of Freesurfer (v6.0) and Civet (v2.1), as well as of other less popular but fresher pipelines (e.g.: Toads-Cruise, ARCTIC, MILXCTE, DiReCT, CLADA) are needed.

This applies for our QCE as well which, in addition, should be tested on datasets hosted in other e-infrastructures. Last but not least, due to the relative small number of some bad labels used in Level-2 (i.e., insula, frontal, occipital, other problems) our QCE might have been affected by unbalance issues, reducing the final performances assessed with the HL and LRAP multi-label metrics.

Future Directions

At the time of writing, multiple efforts are ongoing to capitalize on the significant overlaps and redundancies among different e-infrastructures, as well as to develop seamless and user-transparent interoperability. This is a long-term multinational project, which will eventually lead to the development of a global virtual imaging laboratory. The grand vision is to provide computational neuroscientists with a virtual space, accessible through an ordinary browser, where image data sets and clinical variables, as well as algorithms, pipelines, computational resources, statistical tools, quality-control environments will be transparently accessible irrespective of physical location. A single sign-on system should be used to guarantee user-friendly but still privileged access to non-public resources.

The most mature e-infrastructures could seek convergence towards a worldwide infrastructure that would constitute a global virtual imaging laboratory. This framework is instrumental to the success of ambitious scientific initiatives with high societal impact such as, for example, the prevention of AD before 2020 (PAD 2020 - http://www.pad2020.org).

E-infrastructures will progressively augment their data as well as algorithm portfolios, offering neuroscientists unprecedented power at their fingertips. In this scenario, each new algorithm developed to estimate the cortical thickness (besides those discussed in Chapter 3) should be automatically contrasted vertex-by-vertex via GVF and CPS methodology against the best available tool on the market before being integrated in the e-infrastructures. This would facilitate the job of algorithm developers when deploying and fine-tuning new algorithms, thereby reducing considerably the time needed to be certified as medical devices by recognised boards such as EMA (European Medicines Agency) or FDA (Food and Drug Administration). On the other hand, neuroscientists would be able to use tools to monitor drug efficacy in AD trials which would be more sensitive.

In Chapter 4 we presented the work done to model the QCE using a supervised learning approach. An alternative strategy could have entailed the use of an unsupervised classifier making dimensionality reduction, based for instance on tSNE (t-distributed Stochastic

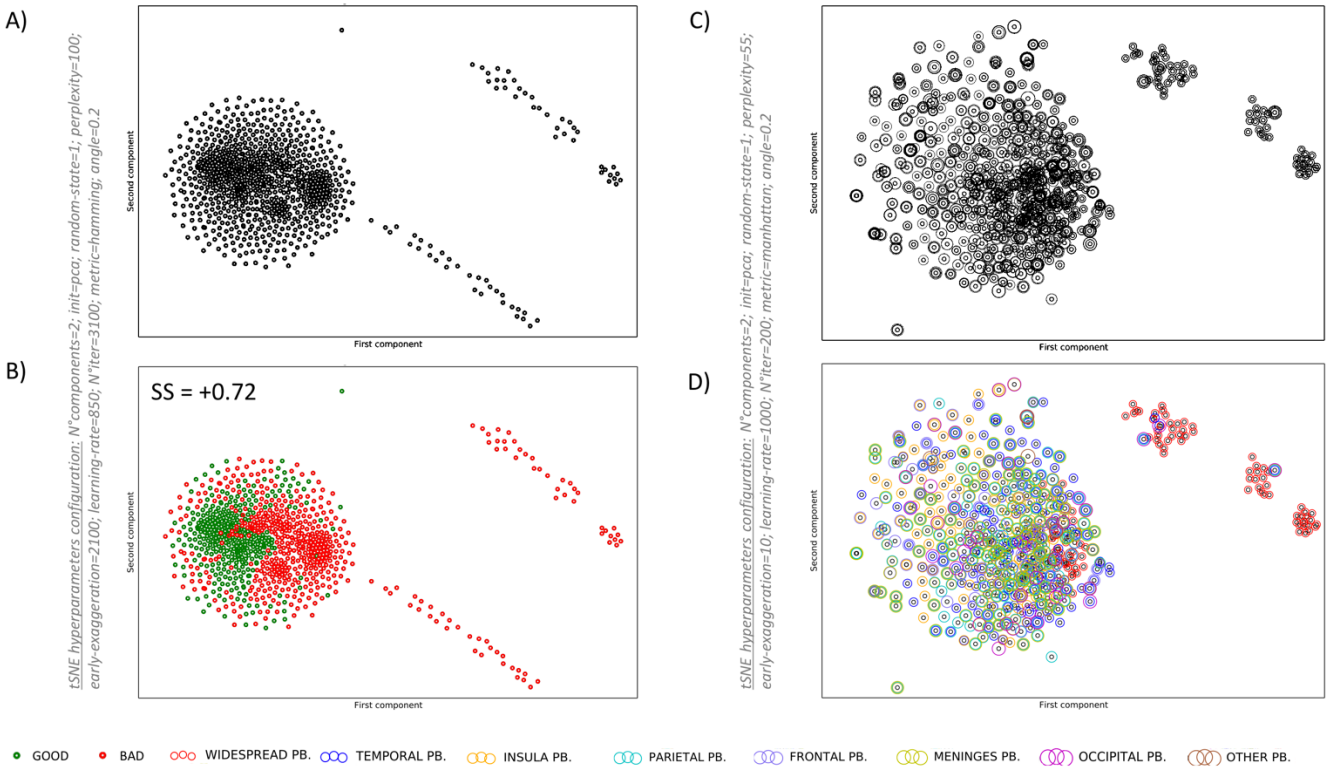


Figure 1 – QCE preliminary results of the tSNE unsupervised dimensionality reduction strategy into a visually plausible 2D space. The tSNE algorithm comprises two main stages. First, tSNE constructs a probability distribution over pairs of high-dimensional objects so that similar objects have a high probability of being picked together, compared with the extremely small probability of dissimilar points. Second, t-SNE defines a similar probability distribution over the points, and it minimizes the Kullback–Leibler divergence between the probabilities of the low-dimensional embedding and the high-dimensional data [5]. Panel A shows 3 main evident clusters after tSNE. Panel B shows the overlay of GOOD and BAD real labels. The global silhouette score (SS) associated with these clusters is +0.72. SS ranges from +1 (best) and -1 (worst). Values near 0 indicate overlapping clusters. Panel C shows 4 main clusters after tSNE. Panel D shows the overlay of the eight bad labels. Although this is a multi-label classification task, we computed the Calinski-Harabasz (CHS) scores just for the eight main atomic labels reported hereinafter to give a glimpse on the quality of cluster segregations: Widespread problems (CHS = 398.87); Temporal (CHS = 7.01); Insula (CHS = 54.20); Parietal (CHS = 29.40); Frontal (CHS = 11.51); Meninges (CHS = 15.93); Occipital (CHS = 0.22); Other problems (CHS = 0.42). A larger CHS value denotes higher compactness.

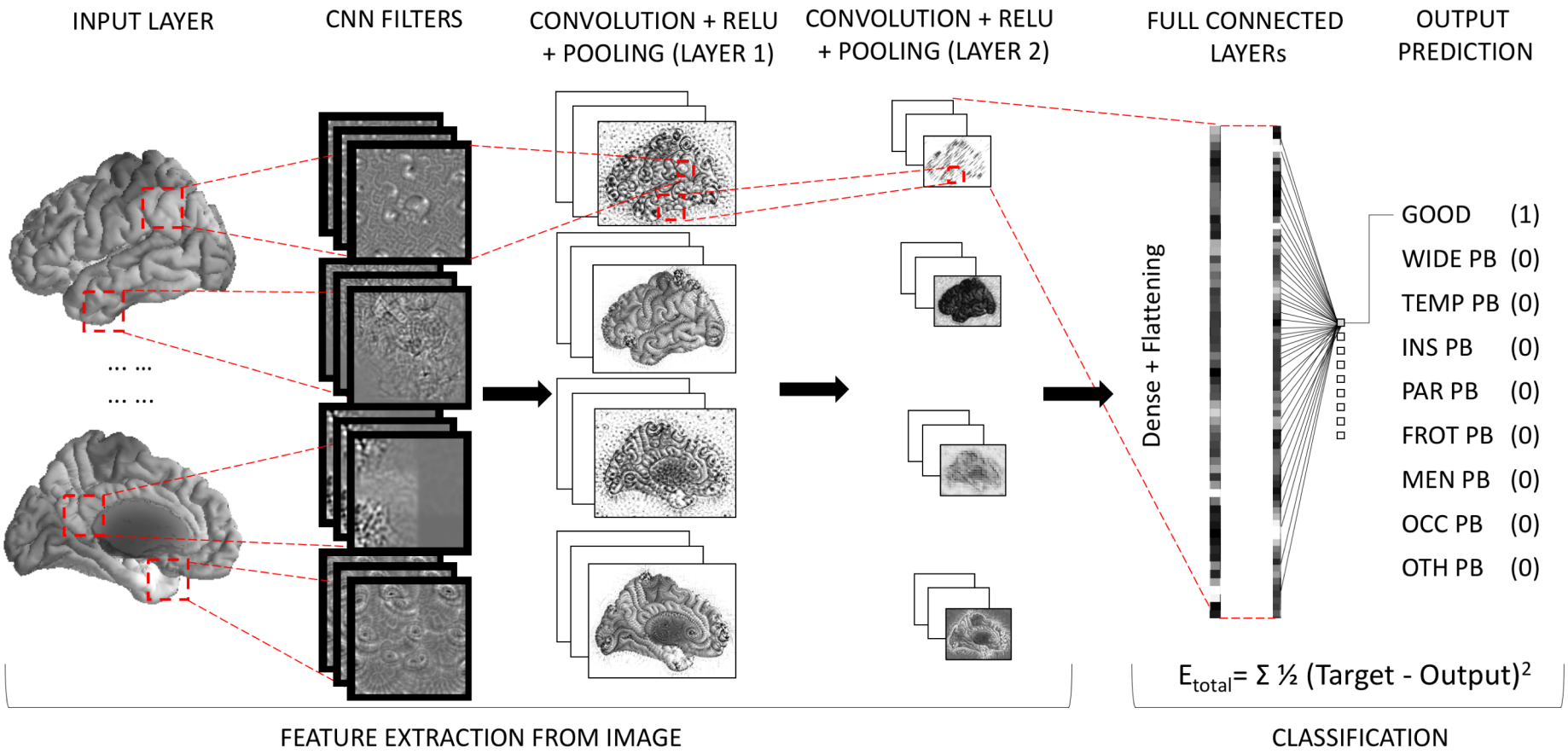


Figure 2 represents a CNN prototype of our QCE. We performed some preliminary tests using both sequential and graph implementations. Our CNN must classify 2D screenshots of a 3D mesh in different (multilabel) categories (i.e., "good", "wide problem", "temporal problem", "insula problem", "parietal problem", "frontal problem", "meninges problem", "occipital problem", "other problem"). As depicted in figure 2, for a mesh provided as input, the CNN correctly assigned the highest probability for the GOOD category among the 9 possible labels. There are four operations in a CNN (i.e., Convolution, Non linearity – RELU, Pooling,

Classification). The primary purpose of Convolution is to extract features from the input image. The features are derived applying different filters. So, a number of feature maps (i.e., sharpened, blurred, etc..) are produced for the same input. In practice, a CNN learns the values of these filters during the training process. The higher number of filters we have, the more image features get extracted, and the better our CNN becomes at recognizing patterns in unseen images. An operation called RELU is used after each Convolution operation. RELU replaces all negative pixel values in the feature map by zero. The purpose of RELU is to introduce non-linearity in our CCN, since most of the real-world pattern we would want our CCN to learn are non-linear. This is also the case for cortical mesh screenshots. The Pooling step reduces the dimensionality of each feature map and retains the most important information. Spatial Pooling can be of different types: Max, Average, Sum, etc. In case of MaxPooling, we defined a spatial neighbourhood (for example, a 3×3 pixel window) and take the largest element from the "RELUed" feature map within that window. In figure 2, we have two sets of Convolution, RELU, and Pooling layers: the 2nd Convolution layer performs convolution on the output of the first Pooling Layer. The output of the 2nd Pooling Layer acts as an input to the Fully Connected Layer. In Keras, we can prepare the input to be fed in the Fully Connected Layer via specific functions called Dense and Flattening. The Fully Connected layer is a traditional Multi Layer Perceptron. The term "Fully Connected" implies that every neuron in the previous layer is connected to every neuron on the next layer. The output from the Convolutional and Pooling layers represent high-level features of the input image. The purpose of the Fully Connected layer is to use these features to classify the input image according to different labels based on the training dataset. For example, the image classification task we set out to perform has 9 possible outputs. For clarity reasons, note that figure 2 does not show connections and weights between the nodes in the Fully Connected layer. The sum of output probabilities from the Fully Connected Layer is 1. Note that in figure 2, since the input image is a GOOD cortical mesh, the target probability obtained from CNN is 1 in the GOOD label and 0 in the other eight (BAD) labels. More concretely, in our example we are testing the CNN using a GOOD 3D mesh, therefore, our target-vector is [1,0,0,0,0,0,0,0,0].

Before testing the CNN on our sample mesh, it is mandatory to train the CNN. To do so, we start by initializing all Weights of the network with random values. As step 2, the CNN takes the first training image as input, goes through the forward propagation step (i.e., Convolution, RELU and Pooling operations along with Forward-propagation in the Fully Connected layer) and outputs probabilities for each class. Let us say that the output probabilities for the first training image (that is GOOD, as the one we would use as testing), are: [0.2,0.4,0.1,0.3,0,0,0,0,0]. Since for the first training example weights are randomly assigned, output probabilities are also random. As step 3, we calculate the total error (E_{total}) at the output layer (summation over all the 9 class-labels). As step 4, we use a Back-propagation to calculate the gradients of the error with respect to all weights in the network, and we use the gradient descent to update all the weights and parameters (in order to minimize the output error). When a similar image is again feed as input, output probabilities could now be [0.7,0.1,0.1,0.1,0,0,0,0,0], which is closer to the target vector [1,0,0,0,0,0,0,0,0]. This means that the network has learnt to classify this particular image correctly by adjusting its weights such that the output error is reduced. Parameters like number of CNN Filters, Filter kernel sizes, number of epochs, CNN Architecture, etc., have all been fixed before step 1 and do not change during the training process. Only the values of

the weights are updated. Finally, we repeat steps 2-4 with all the images in the training set. At the end, all the weights of the CNN are optimized to correctly classify images from the training set.

When a new (unseen) image is fed to the CNN, the network would go through the forward propagation step and output a probability for each class (in this case, the output probabilities are calculated using the weights previously optimized to correctly classify all the training examples).

If the training set is large enough, the CNN will (hopefully) generalize well to new images and classify them into correct categories. In our case, the GOOD input image reported in figure2 has been correctly classified as GOOD; however, the overall performance we registered of our prototype could be improved following the so called "aggressive data augmentation" and hyper-parameter optimizations. Below we report the overall performance using 1'069 training images and 268 testing images in 100 epochs. Epoch #1 (starting point): TRAIN_ACCURACY = 0.52, TEST_ACCURACY = 0.51; Epoch #100 (ending point): TRAIN_ACCURACY=0.70, TEST_ACCURACY = 0.68.

Neighbor Embedding) in association with a Nearest Neighbour classification algorithms (figure 1), or a deep-learning approach (via Theano[1] API & Keras[2] libraries) and convolutional neural networks (CNN) [3] (figure 2).

In the first case, the algorithm would discover and present the interesting structure in the data. This could be useful to model the underlying data structures and learn more about the intrinsic features of the data, a typical semi-supervised or unsupervised machine learning task. Indeed, the same 3D meshes' descriptors used in the supervised learning approach could be employed here, perhaps adding new promising features.

In the second case, we would need to input screenshots of 3D meshes to QCE, saving a lot of time in the calculation of numeric features. To reduce computational time, the CNN would benefit of Graphics Processing Unit (GPU) power as well. In the deep-learning stage, we should consider to heavily expand the datasets through the so-called "aggressive data augmentation" strategy in order to let the CNN generalize and learn sufficiently well how to judge mesh artefacts.

In the next future we plan to extend our QCE to other imaging modalities (e.g., DTI, T2-FLAIR, R-fMRI) and pipelines, to perform automatic QC on biomarkers of interest for the neuroimaging field [6] (e.g., hippocampus, amygdala, basal ganglia, and other subcortical ROI segmentations, white matter lesion segmentations, white matter tract delineations).

[1] http://deeplearning.net/software/theano/library/index.html
[2] https://keras.io/

REFERENCES

1. Terrence J Sejnowski, Patricia S Churchland, and J Anthony Movshon. Putting big data to good use in neuroscience. Nat Neurosci. 2014 Nov; 17(11): 1440–1441.

2. Toga AW, Foster I, Kesselman C, Madduri R, Chard K, Deutsch EW, Price ND, Glusman G, Heavner BD, Dinov ID, Ames J, Van Horn J, Kramer R, Hood L. Big biomedical data as the key resource for discovery science. J Am Med Inform Assoc. 2015 Nov;22(6):1126-31. doi: 10.1093/jamia/ocv077. Epub 2015 Jul 21.

3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436-44. doi: 10.1038/nature14539.

4. Mwangi B, Soares JC, Hasan KM. Visualization and unsupervised predictive clustering of high-dimensional multimodal neuroimaging data. J Neurosci Methods. 2014 Oct 30;236:19-25. doi: 10.1016/j.jneumeth.2014.08.001. Epub 2014 Aug 10.

5. Van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

6. Pini L, Pievani M, Bocchetta M, Altomare D, Bosco P, Cavedo E, Galluzzi S, Marizzoni M, Frisoni GB. Brain atrophy in Alzheimer's Disease and aging. Ageing Res Rev. 2016 Sep;30:25-48. doi: 10.1016/j.arr.2016.01.002. Epub 2016 Jan 28.

Addendum

List of publications

&

Résumé conclusif

LIST OF PUBLICATIONS DURING THE PhD PERIOD

Work described in this PhD thesis

- *Chapter 4:*

  Redolfi A, Fischer C, Orlandi D, Duchesnay É, Operto G, Rivière D, Frisoni GB, Mangin JF, neuGRID and CATI Consortia, for the Alzheimer's Disease Neuroimaging Initiative. QCE (Quality Control Environment): A machine learning tool for automatic classification of cortical meshes. NeuroImage. 2017 - *submitted*.

- *Chapter3:*

  Redolfi A, Manset D, Barkhof F, Wahlund LO, Glatard T, Mangin JF, Frisoni GB; neuGRID Consortium, for the Alzheimer's Disease Neuroimaging Initiative. Head-to-head comparison of two popular cortical thickness extraction algorithms: a cross-sectional and longitudinal study.PLoS One. 2015 Mar 17;10(3):e0117692. doi: 10.1371/journal.pone.0117692. eCollection 2015.

- *Chapter 2:*

  Redolfi A, Bosco P, Manset D, Frisoni GB; neuGRID consortium. Brain investigation and brain conceptualization. Funct Neurol. 2013 Jul-Sep;28(3):175-90. doi: 10.11138/FNeur/2013.28.3.175.

Others

- Bosco P, **Redolfi A**, Bocchetta M, Ferrari C, Mega A, Galluzzi S, Austin M, Chincarini A, Collins DL, Duchesne S, Maréchal B, Roche A, Sensi F, Wolz R, Alegret M, Assal F, Balasa M, Bastin C, Bougea A, Emek-Savaş DD, Engelborghs S, Grimmer T, Grosu G, Kramberger MG, Lawlor B, Mandic Stojmenovic G, Marinescu M, Mecocci P, Molinuevo JL, Morais R, Niemantsverdriet E, Nobili F, Ntovas K, O'Dwyer S, Paraskevas GP, Pelini L, Picco A, Salmon E, Santana I, Sotolongo-Grau O, Spiru L, Stefanova E, Popovic KS, Tsolaki M, Yener GG, Zekry D, Frisoni GB. The impact of automated hippocampal volumetry on diagnostic confidence in patients with suspected Alzheimer's disease: An EADC study. Alzheimers Dement. 2017 Mar 3. pii: S1552-5260(17)30045-6. doi: 10.1016/j.jalz.2017.01.019. [Epub ahead of print].
- Cover KS, van Schijndel RA, Versteeg A, Leung KK, Mulder ER, Jong RA, Visser PJ, **Redolfi A**, Revillard J, Grenier B, Manset D, Damangir S, Bosco P, Vrenken H, van Dijk BW, Frisoni GB, Barkhof F; Alzheimer's Disease Neuroimaging Initiative,

neuGRID. Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease. Psychiatry Res. 2016 Jun 30;252:26-35. doi: 10.1016/j.pscychresns.2016.04.006. Epub 2016 May 11.

- Maglietta R, Amoroso N, Boccardi M, Bruno S, Chincarini A, Frisoni GB, Inglese P, **Redolfi A**, Tangaro S, Tateo A, Bellotti R; Alzheimers Disease Neuroimaging Initiative.Automated hippocampal segmentation in 3D MRI using random undersampling with boosting algorithm. Pattern Anal Appl. 2016;19:579-591. Epub 2015 Jul 9.
- Chincarini A, Sensi F, Rei L, Gemme G, Squarcia S, Longo R, Brun F, Tangaro S, Bellotti R, Amoroso N, Bocchetta M, **Redolfi A**, Bosco P, Boccardi M, Frisoni GB, Nobili F; Alzheimer's Disease Neuroimaging Initiative. Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease. Neuroimage. 2016 Jan 15;125:834-47. doi: 10.1016/j.neuroimage.2015.10.065. Epub 2015 Oct 26.
- Inglese P, Amoroso N, Boccardi M, Bocchetta M, Bruno S, Chincarini A, Errico R, Frisoni GB, Maglietta R, **Redolfi A**, Sensi F, Tangaro S, Tateo A, Bellotti R; Alzheimer's Disease Neuroimaging Initiative. Multiple RF classifier for the hippocampus segmentation: Method and validation on EADC-ADNI Harmonized Hippocampal Protocol. Phys Med. 2015 Dec;31(8):1085-91. doi: 10.1016/j.ejmp.2015.08.003. Epub 2015 Oct 21.
- Boccardi M, Bocchetta M, Morency FC, Collins DL, Nishikawa M, Ganzola R, Grothe MJ, Wolf D, **Redolfi A**, Pievani M, Antelmi L, Fellgiebel A, Matsuda H, Teipel S, Duchesne S, Jack CR Jr, Frisoni GB; EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimers Dement. 2015 Feb;11(2):175-83. doi: 10.1016/j.jalz.2014.12.002. Epub 2015 Jan 20.
- Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, DeCarli C, deToledo-Morrell L, Firbank M, Ganzola R, Gerritsen L, Henneman W, Killiany RJ, Malykhin N, Pasqualetti P, Pruessner JC, **Redolfi A**, Robitaille N, Soininen H, Tolomeo D, Wang L, Watson C, Wolf H, Duvernoy H, Duchesne S, Jack CR Jr, Frisoni GB; EADC-ADNI Working Group on the Harmonized Protocol for Manual

Hippocampal Segmentation. Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimers Dement. 2015 Feb;11(2):126-38. doi: 10.1016/j.jalz.2014.02.009. Epub 2014 Aug 15.

- Cover KS, van Schijndel RA, Popescu V, van Dijk BW, **Redolfi A**, Knol DL, Frisoni GB, Barkhof F, Vrenken H; neuGRID.; Alzheimer′s Disease Neuroimaging Initiative. The SIENA/FSL whole brain atrophy algorithm is no more reproducible at 3T than 1.5 T for Alzheimer's disease. Psychiatry Res. 2014 Oct 30;224(1):14-21. doi: 10.1016/j.pscychresns.2014.07.002. Epub 2014 Jul 14.
- Tangaro S, Amoroso N, Boccardi M, Bruno S, Chincarini A, Ferraro G, Frisoni GB, Maglietta R, **Redolfi A**, Rei L, Tateo A, Bellotti R; Alzheimers Disease Neuroimaging Initiative. Automated voxel-by-voxel tissue classification for hippocampal segmentation: methods and validation. Phys Med. 2014 Dec;30(8):878-87. doi: 10.1016/j.ejmp.2014.06.044. Epub 2014 Jul 11.
- Cavedo E, **Redolfi A**, Angeloni F, Babiloni C, Lizio R, Chiapparini L, Bruzzone MG, Aquino D, Sabatini U, Alesiani M, Cherubini A, Salvatore E, Soricelli A, Vernieri F, Scrascia F, Sinforiani E, Chiarati P, Bastianello S, Montella P, Corbo D, Tedeschi G, Marino S, Baglieri A, De Salvo S, Carducci F, Quattrocchi CC, Cobelli M, Frisoni GB. J Alzheimers Dis. The Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI): validation of structural MR imaging. 2014;40(4):941-52. doi: 10.3233/JAD-132666.
- Boccardi M, Bocchetta M, Ganzola R, Robitaille N, **Redolfi A**, Duchesne S, Jack CR Jr, Frisoni GB; EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimers Dement. 2015 Feb;11(2):184-94. doi: 10.1016/j.jalz.2013.03.001. Epub 2013 May 21.
- Babiloni C, Carducci F, Lizio R, Vecchio F, Baglieri A, Bernardini S, Cavedo E, Bozzao A, Buttinelli C, Esposito F, Giubilei F, Guizzaro A, Marino S, Montella P, Quattrocchi CC, **Redolfi A**, Soricelli A, Tedeschi G, Ferri R, Rossi-Fedele G, Ursini F, Scrascia F, Vernieri F, Pedersen TJ, Hardemark HG, Rossini PM, Frisoni GB. Resting state cortical electroencephalographic rhythms are related to gray matter volume in subjects with mild cognitive impairment and Alzheimer's disease. Hum Brain Mapp. 2013 Jun;34(6):1427-46. doi: 10.1002/hbm.22005. Epub 2012 Feb 14.

RÉSUMÉ CONCLUSIF

L'âge mur de l'imagerie cérébrale *in vivo* génère aujourd'hui un déluge d'informations numériques sur le cerveau humain. Des études de recherche en imagerie cérébrale sont effectuées pour surveiller le cerveau en action ou au repos, pour examiner comment il est construit, comment les connections se font en son sein, et ce qui se passe lorsque les choses tournent mal. La croissance fantastique, la disponibilité et l'accessibilité des données de l'imagerie par résonance magnétique (IRM) issues de personnes souffrant de troubles neurodégénératifs, par exemple la maladie d'Alzheimer (MA), ont stimulé le développement de nouveaux environnements informatiques tels que les infrastructures électroniques. Ces plates-formes, accessibles en ligne via les navigateurs Web communs, offrent aux scientifiques de grandes bases de données d'images, des pipelines sophistiqués pour l'analyse des images, des ressources informatiques puissantes, des algorithmes pour une visualisation tridimensionnelle et des outils statistiques. Toutes ces composantes contribuent au développement de l'ère actuelle des « Big Data ».

Pour comprendre le changement récent du paradigme neuroscientifique de la recherche dans l'ère des « Big Data » et pour donner un aperçu des expériences sophistiquées qu'un neuroscientifique peut effectuer aujourd'hui, nous avons supposé **dans le Chapitre 2** la construction d'un hypermodèle multi-échelle et multimodal basé sur l'inférence bayésienne et le modèle hiérarchique à effets mixtes.

En effet, le cerveau d'un patient atteint de la maladie d'Alzheimer subit des changements qui débutent plusieurs années avant le développement des premiers symptômes cliniques. En outre, les changements cérébraux chez les patients atteints de la maladie d'Alzheimer se produisent à différents niveaux et pour différentes raisons : au niveau moléculaire, les changements sont dus au dépôt amyloïde ; au niveau cellulaire, à la perte de synapses neuronales, et au niveau tissulaire, à une perte de connectivité. Ils causent tous une atrophie étendue de l'ensemble de l'organe cérébral.

Les neurosciences modernes ont permis d'approfondir tous les échelles de l'organisation du cerveau, des gènes à la cognition. Cependant, les neurosciences se heurtent à la nécessité impérieuse de mieux organiser les modèles et de les adapter aux différentes échelles, afin de capter, grâce à des règles mathématiques solides, la « mécanique profonde » du cerveau.

Mon hypermodèle pourrait être considéré comme un d'ordre supérieur de la progression de la maladie d'Alzheimer. Malheureusement, cet hypermodèle du cerveau est encore dépendant

d'un très grand nombre de variables. Jusqu'à présent, peu de neuroscientifiques ont pu concrètement réaliser un tel hypermodèle, principalement en raison du manque de données et d'un environnement *ad hoc* disposant d'une puissance suffisante.

Les hypermodèles du cerveau doivent exploiter les infrastructures électroniques avancées en les utilisant comme moteur de calcul pour surmonter leurs besoins arithmétiques élevés. Heureusement, il a été entrepris d'organiser de manière remarquable l'accessibilité et la disponibilité des données d'imagerie et de non-imagerie des personnes touchées par la maladie d'Alzheimer dans de nombreuses infrastructures électroniques (à savoir : Centre pour l'Acquisition et le Traitement de l'Image – CATI ; Virtual Imaging Platform – VIP ; European Medical Information Framework – EMIF ; Global Alzheimer's Association Interactive Network – GAAIN ; Laboratory of Neuroimaging – LONI ; Canadian Brain Imaging Research Platform – CBRAIN ; neuGRID). Ces infrastructures électroniques permettent d'effectuer des expériences en imagerie médicale à l'aide de ressources informatiques dédiées telles que des grilles, des systèmes de calcul de haute performance (HPC) et des clouds publics ou privés. La quantité de ressources informatiques, de données et d'algorithmes à manipuler supporte l'idée que les infrastructures électroniques sont aujourd'hui les systèmes les mieux équipés pour soutenir la création de modèles cérébraux avancés. Ainsi, un neuroscientifique serait juste à un clic de tout ce dont il a besoin pour commencer ses simulations.

Plus récemment ont été initiées en Europe et aux États-Unis des initiatives richissimes basées sur l'intégration des infrastructures électroniques existantes ou sur la création de nouvelles plates-formes électroniques, visant à modéliser l'ensemble du cerveau humain (à savoir : le Humain Brain Project [HBP] pour l'Europe et l'initiative Brain Activity Map [BAM] pour les États-Unis) dans le but de réduire dans les cinq prochaines années le fardeau des maladies du cerveau, grâce à des approches complètes axées sur les données.

Ce moment historique sera certainement destiné à laisser une grande empreinte chez les neuroscientifiques et dans la manière dont nous menons les (neuro)sciences.

Dans l'attente d'un hypermodèle multimodal et multi-échelle entièrement validé pour la maladie d'Alzheimer, dans le **Chapitre 3**, nous avons effectué une comparaison directe entre les deux méthodes les plus connues pour estimer l'épaisseur corticale, à savoir : Freesurfer (FS) and Civet-CLASP (CV).

L'amincissement cortical est considéré comme un marqueur en trois dimensions prometteur de la neurodégénérescence, un marqueur putatif de la progression de la maladie et un biomarqueur de substitution raisonnable dans les essais cliniques.

Nous avons analysé les données ADNI provenant de 69 sujets âgés contrôles (CTR), 37 sujets atteints de déficiences cognitives légères stables (DCLs), 27 sujets atteints de déficiences cognitives légères progressives (DCLp) et 52 sujets scannées après le diagnostic de la maladie d'Alzheimer. Les images utilisées ont été acquises 12 mois et 24 mois après l'inclusion dans l'étude.

L'expérience a été réalisée à l'aide de deux infrastructures électroniques décrites précédemment (à savoir : VIP et neuGRID). Cette étude a représenté des expériences très lourdes d'un point de vue informatique et de stockage. Au total, nous avons analysé 1110 scans volumétriques de grande dimension pour chaque pipeline, générant 0,55 To de données dérivées. Grâce à la parallélisation des travaux, nous avons réalisé cette tâche en 10 semaines plutôt qu'en 5 ans sur un ordinateur monoprocesseur.

Notamment, cette étude a représenté la première tentative pour examiner les forces et les faiblesses mutuelles de Civet et de Freesurfer, dans un vrai défi en face à face, en considérant des sujets sur l'ensemble du spectre de la maladie et à l'échelle de précision du voxel. Dans la littérature, seules des méthodes de validation basées sur des fantômes ont été décrites, mais cette approche ne peut pas prendre en compte l'ensemble des aspects des données réelles.

Civet et Freesurfer sont des pipelines très différents, caractérisés par des procédures algorithmiques et des conventions spécifiques. Cela rend la comparaison des résultats difficile. Ce problème a été résolu en développant une approche innovante. Plus précisément, les flux de vecteurs gradients 3D (GVF) et la recherche du point le plus proche (CPS) ont été appliqués pour assurer une comparaison fiable des différents sommets des maillages. En effet, les maillages d'origine de Freesurfer et de Civet étaient caractérisés par une morphométrie et une topographie complètement différentes. Cependant, grâce à la comparaison directe par algorithme croisé vertex par vertex, les différences entre les deux pipelines, à la fois pour l'analyse transversale et longitudinale, ont été quantifiées de manière analytique grâce à un modèle hybride dédié généré à partir des modèles respectifs de Freesurfer et de Civet.

Les performances de Civet et de Freesurfer ont varié selon le stade de la maladie, soulignant qu'aucun des deux algorithmes ne pouvait être considéré systématiquement comme le meilleur. Dans le cadre d'une analyse transversale, Civet a différé de manière significative ($p<0,05$) de Freesurfer dans les grandes régions frontales, pariétales, temporales et occipitales. Dans une

analyse discriminante (CTR vs DCLp, CTR vs MA) avec des régions d'intérêt corticales ayant une taille d'effet supérieure à 0,8, les deux pipelines n'ont montré aucune différence significative dans une étude par courbe ROC.
Longitudinalement, Freesurfer a différé de manière significative de Civet ($p < 0,05$) dans le gyrus supramarginal, le gyrus temporal et dans le cortex occipital latéral. Dans une analyse discriminante (DCLp au départ VS DCLp à moins de 24 mois) avec des régions d'intérêt ayant une taille d'effet supérieure à 0,6, les deux pipelines n'ont montré aucune différence significative dans le cadre ROC (Receiver Operating Characteristic).
Civet est apparu légèrement plus sensible au modèle atrophique typique de la maladie d'Alzheimer dans les DLCp, mais les deux pipelines caractérisent avec précision la topographie de l'amincissement cortical au stade de la démence. En définitive, avoir la possibilité de définir des pipelines fiables permettrait aux médecins d'identifier au moment opportun les patients destinés à connaitre une évolution rapide de leur MA, ce qui sera très bénéfique lorsque des traitements modifiant l'évolution de la maladie deviendront disponibles.

Pour extraire des informations scientifiquement pertinentes, un neuroscientifique ne peut plus inspecter manuellement les milliers d'images cérébrales qui sont traitées de manière efficace par les infrastructures électroniques. Malheureusement, jusqu'à présent, les méthodes de contrôle qualité (CQ) et les procédures de validation des résultats n'ont été effectuées que par des scientifiques experts, et seule une inspection visuelle a été considérée. Ce contrôle qualité était impératif, car aucun algorithme ou pipeline disponible aujourd'hui dans le domaine de la neuroimagerie n'est capable de générer des résultats 100% corrects.
Par conséquent, l'étape suivante de cette thèse était de développer une méthode capable de réduire le temps nécessaire pour effectuer un CQ précis sur l'énorme quantité de données générées par les infrastructures électroniques, grâce à des algorithmes avancés de segmentation de la surface corticale, telles que Freesurfer et Civet. Dans le **Chapitre 4**, nous avons conçu un environnement automatique pour le contrôle qualité automatique (ECQ) qui vise à créer une nouvelle base pour l'évaluation des résultats des pipelines.
Nous avons traité et collecté 1582 résultats Freesurfer et 1692 résultats Civet, qui étaient issus de 5 ensembles de données liés à plusieurs cohortes et un grand nombre de scanners, à savoir : ADNI, ARWIBO, EDSD, OASIS, et PharmaCOG. L'ECQ a été validé par rapport à une évaluation visuelle, ce qui est la référence pour les experts. Dans cette étude, les labels pour identifier les artefacts des maillages ont été définis à travers un panel Delphi basé sur des

preuves, par trois experts qui ont établi un consensus basé sur leur expérience personnelle. Les évaluateurs étaient des neuroscientifiques ayant plus de 5 ans d'expérience dans le domaine de la neuroimagerie.

Un nouvel ensemble de données de référence pour le CQ a été construit et rendu public dans l'infrastructure électronique neuGRID pour la communauté de la neuroimagerie. Il contient les maillages corticaux 3D de Freesurfer et de Civet, avec des labels individuels et multiples.

L'ECQ a été conçue comme un classifieur multiniveau. Au niveau 1, une forêt d'arbres décisionnels (RF) permet une classification des « bons » et des « mauvais » maillages corticaux. Au niveau 2, un classifieur d'ensemble souple composé par un séparateur à vaste marge (SVM), avec une RF, permet un étiquetage multilabel de 8 artefacts (à savoir : « problèmes généralisés » ; « problèmes temporaux » ; « problèmes d'insula » ; « problèmes pariétaux » ; « problèmes frontaux » ; « problèmes de méninges » ; « problèmes occipitaux » ; « autres problèmes ») sur les mauvais maillages.

Pour notre approche par apprentissage automatique (AA) supervisé, nous avons extrait un vecteur de caractéristiques à partir des maillages 3D. Nous avons effectué une phase de prétraitement, principalement caractérisée par le recalage des maillages dans le référentiel de Talairach et des conversions de format de données.

Plusieurs fonctions basées sur des descripteurs de biomarqueurs 3D dérivés (à savoir : test de rapport de vraisemblance de l'épaisseur corticale, descripteur de forme des sillons, etc.), des propriétés géométriques 2D (à savoir : approche fondée sur l'image 2D), ou les mesures de distorsion 3D (approche basée sur les modèles de maillage) ont été extraites des maillages de Freesurfer et Civet.

Les résultats ont montré que l'ECQ se comportait de la même manière dans Freesurfer et Civet. Des courbes ROC ont été utilisées pour évaluer la puissance discriminative de la classification des « bons » et des « mauvais » dans les deux pipelines (AUC^{FS}_{Gauche}=0,96 ; AUC^{FS}_{Droit} =0,94 ; AUC^{CV}_{Gauche}=0,96 ; AUC^{CV}_{Droit} =0,91). L'ECQ a semblé aussi solide pour les deux pipelines au niveau 1.

Au niveau 2, la perte de Hamming (HL) de l'ECQ a été systématiquement meilleure dans Freesurfer, le score de la précision moyenne de la classification des labels (LRAP) était légèrement moins bon dans Civet.

Notre objectif n'était pas une tâche facile car, pour remplacer le travail d'un expert, l'ECQ doit disposer : (i) d'une sensibilité élevée et (ii) d'une erreur de type II suffisamment faible.

Selon nos expériences, au niveau 1, l'ECQ a produit de brillants résultats, avec une précision élevée et une sensibilité acceptable, et correspondant à ceux obtenus par des experts normalement formés. Au niveau 2, l'ECQ a été suffisamment performante pendant la phase de validation multilabel, bien qu'elle pâtisse de labels manquants ou incorrects.
Par rapport à l'évaluation visuelle, l'ECQ représente une solution idéale en termes d'évolutivité. Elle peut classer des milliers de maillages corticaux 3D plus efficacement qu'un expert. Cette étape manuelle est d'ailleurs inévitablement sujette à des erreurs subjectives lors de l'inspection visuelle.
Il est maintenant possible de détecter des erreurs dans les maillages corticaux 3D d'une manière objective et automatique. L'ECQ représente une occasion unique de traiter plus rapidement les données, ce qui permet aux neuroscientifiques de passer leur temps précieux à effectuer des analyses de données plutôt que des contrôles qualité fastidieux.

Dans l'ensemble, les études rapportées ont démontré plusieurs points forts en présentant des avantages directs sur : (i) l'utilisation d'une puissante infrastructure virtuelle efficacement interfacée avec d'autres, à savoir, neuGRID ; (ii) des ressources informatiques élevées, (iii) des ensembles de données larges et hétérogènes (qui ont permis une généralisation accrue de nos résultats), (iv) des algorithmes puissants pour l'analyse des images.
Dans cette thèse, nous avons prouvé l'utilité et amélioré le service des infrastructures dématérialisées. Les produits et les outils qui ont été développés aideront à progresser et à développer davantage les services des infrastructures dématérialisées et la façon dont nous pourrons mener les futures études de neuroimagerie.

Titre : E-infrastructure, segmentation du cortex, environnement de contrôle qualité : un fil rouge pour les neuroscientifiques

Mots clés : e-infrastructure, apprentissage automatique, maladie d'Alzheimer, maillage 3D, épaisseur corticale, contrôle de qualité.

Résumé : Les neurosciences sont entrées dans l'ère des « big data ». Les ordinateurs de bureau individuels ne sont plus adaptés à l'analyse des téraoctets et potentiellement des pétaoctets qu'impliquent les images cérébrales. Pour combler le gouffre qui existe entre la taille des données et les possibilités standard d'extraction des informations, on développe actuellement des infrastructures virtuelles en Amérique du Nord, au Canada et également en Europe. Ces infrastructures dématérialisées permettent d'effectuer des expériences en imagerie médicale à l'aide de ressources informatiques dédiées telles que des grilles, des systèmes de calcul haute performance (HPC) et des clouds publics ou privés. Les infrastructures virtuelles sont aujourd'hui les systèmes les plus avancés et les mieux équipés pour soutenir la création de modèles multimodaux et multi-échelles avancés du cerveau atteint par la maladie d'Alzheimer (Chapitre 2) ou pour valider des biomarqueurs d'imagerie prometteurs, tels que l'épaisseur corticale, grâce à des pipelines sophistiqués (Chapitre 3). En effet, les analyses d'imagerie, telles que celles décrites dans les Chapitres 2 et 3, multiplient de manière exponentielle la quantité de données post-traitées qui atteignent, à la fin, des téraoctets de résultats pour une seule étude. Afin de faire face à l'énorme quantité de données de post-traitement générées par les infrastructures électroniques, un environnement de contrôle qualité automatique (ECQ) des maillages de la surface corticale (Chapitre 4) a été proposé. L'ECQ est un classifieur par apprentissage automatique (AA) avec une approche par apprentissage supervisé basée sur les forêts d'arbres décisionnels (RF) et des estimations par séparateurs à vaste marge (SVM). Compte tenu de son évolutivité et de son efficacité, l'ECQ s'inscrit bien dans les infrastructures électroniques en cours de développement, où ce type de service de vérification élémentaire manque toujours. L'ECQ représente une occasion unique de traiter les données plus facilement et plus rapidement, ce qui permettra aux neuroscientifiques de passer leur temps précieux à effectuer des analyses de données au lieu de le dépenser dans des tâches manuelles et laborieuses de contrôle qualité.

Title : E-infrastructure, cortical mesh segmentation, quality control environment: a red thread for neuroscientists

Keywords : e-infrastructure, machine learning, Alzheimer's disease, 3D mesh, cortical thickness, quality control.

Abstract : Neuroscience entered the "big data" era. Individual desktop computers are no longer suitable to analyse terabyte, and potentially petabytes, of brain images. To fill in the gap between data acquisition and information extraction, e-infrastructures are being developing in North America, Canada, and Europe. E-infrastructures allow neuroscientists to conduct neuroimaging experiments using dedicated computational resources such as grids, high-performance computing (HPC) systems, and public/private clouds. Today, e-infrastructures are the most advanced and the best equipped systems to support the creation of advanced multimodal and multiscale models of the Alzheimer's disease (AD) brain (Chapter 2) or to validate promising imaging biomarkers with sophisticated pipelines, as for cortical thickness, (Chapter 3). Indeed, imaging analyses such as those described in Chapter 2 and 3 expand the amount of post-processed data per single study. In order to cope with the huge amount of post-processing data generated via e-infrastructures, an automatic quality control environment (QCE) of the cortical delineation algorithms is proposed (Chapter 4). QCE is a machine learning (ML) classifier with a supervised learning approach based on Random Forest (RF) and Support Vector Machine (SVM) estimators. Given its scalability and efficacy, QCE fits well in the e-infrastructures under development, where this kind of sanity check service is still lacking. QCE represents a unique opportunity to process data more easily and quickly, allowing neuroscientists to spend their valuable time do data analysis instead of using their resources in manual quality control work.