



Development and evaluation of AI methods for multi-centric MRI harmonization: application to pediatric brain studies

Stenzel Cackowski

► To cite this version:

Stenzel Cackowski. Development and evaluation of AI methods for multi-centric MRI harmonization : application to pediatric brain studies. Physics [physics]. Université Grenoble Alpes [2020-..], 2022. English. NNT : 2022GRALY067 . tel-04002441

HAL Id: tel-04002441

<https://theses.hal.science/tel-04002441>

Submitted on 23 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : PHYS - Physique

Spécialité : Physique pour les Sciences du Vivant

Unité de recherche : Grenoble Institut des Neurosciences

Développement et évaluation de méthodes d'IA pour l'harmonisation d'IRM multi-centrique : application aux études du cerveau pédiatrique

Development and evaluation of AI methods for multi-centric MRI harmonization : application to pediatric brain studies

Présentée par :

Stenzel CACKOWSKI

Direction de thèse :

Emmanuel BARBIER

DIRECTEUR DE RECHERCHE, Université Grenoble Alpes

Directeur de thèse

Michel DOJAT

DIRECTEUR DE RECHERCHE, Université Grenoble Alpes

Co-directeur de thèse

Rapporteurs :

MERITXELL BACH CUADRA

Professeur assistant, Université de Lausanne

OLIVIER COLLIOT

Directeur de recherche, CNRS DELEGATION PARIS CENTRE

Thèse soutenue publiquement le **4 novembre 2022**, devant le jury composé de :

MERITXELL BACH CUADRA

Professeur assistant, Université de Lausanne

Rapporteure

OLIVIER COLLIOT

Directeur de recherche, CNRS DELEGATION PARIS CENTRE

Rapporteur

JULIEN THEVENON

Professeur des Univ. - Praticien hosp., UNIVERSITE GRENOBLE ALPES

Examineur

Président du jury

NATHALIE BODDAERT

Professeur des Univ. - Praticien hosp., UNIVERSITE DE PARIS-CITE

Examinatrice

Invités :

THOMAS CHRISTEN

Chargé de recherche, GrenUNIVERSITE GRENOBLE ALPES

Abstract

This PhD project was initiated by the French national project DEFIDIAG, focusing on Intellectual Disorders (ID) diagnosis among children. It was supported by MIAI, the Grenoble Multidisciplinary Institute in Artificial Intelligence. In a context involving a large multi-centric database gathering MRI data from children with a large age range, we addressed several challenges and proposed new image processing tools in order to pave the way for further radiomic analyses.

The first concern about the DEFIDIAG project was the population age range and its consequences on further analyses. After a bibliographic review about brain development, we assessed the importance of biological covariates such as age, sex or ID on brain trajectories. Moreover, we were able to set up a coherent pre-processing workflow in order to homogenize the data while preserving the biological variations induced by age. This was mainly done by using an open access age-specific templates database, representing averaged brains at 6 months intervals.

The second concern was MR data homogeneity between the multiple sites of acquisition. The process of removing site or scanner related noises while preserving biological information in scans is called data ‘Harmonization’ and has become our main point of focus.

In the first step, we reviewed several data harmonization solutions that have been proposed in the literature with a special focus on nonlinear analyses using Artificial Intelligence tools.

Then, we compared a recent deep learning-based solution, ‘cycleGAN’, to ComBAT, a statistical solution considered as a reference in the domain. This study on real and synthetic data led to interesting results that emphasized the great potential of deep learning-based approaches. On the other hand, cycleGAN’s architecture was limited to ‘two site harmonization’ only and was thus difficult to use in most multi-centric studies scenarios. Other practical limitations were identified and we concluded that no existing harmonization solution was suited for our needs.

Therefore, we proposed an original DL harmonization solution called ImUnity and proposed complementary experiments to validate our approach. Harmonization on traveling subjects led to state-of-the-art harmonization results. Using two classification tasks (1: site; 2: status), we showed that ImUnity was able to remove site effects and preserve biological information. Additionally, it was shown that the solution could generalize its training to data coming from unseen sites.

In a final study, we estimated the need and impact of MR harmonization in brain developmental analysis. This was done through multiple approaches that included: 1) An investigation on brain development evolution with age compared to literature results. 2) The introduction of a metric to compare trends before and after harmonization with reference from literature 3) Statistical tests using a linear mixed model approach to better estimate harmonization impact on features (biological or not). 4) A group analysis effect ran to highlight brain surface areas mostly concerned by harmonization. Overall, we observed significant improvement of ROIs volume evolutions after harmonization with similar performance for healthy volunteers and autistic patients. The impact on cortical thickness evolution was less significant and negative impacts of harmonization were even observed in some ROIs. Further investigations are thus required in order to fully understand the effect of harmonization on this type of metrics.

In conclusion, our investigations on MRI data coming from children from a large age

range as well as multiple acquisition sites have led to the development of innovative image processing tools. Even if new features can be added (3D approach, multiple contrasts analyses, etc.) and further validations can be performed (impact of geometrical distortions, study on larger databases, etc.), ImUnity seems to be able to provide high quality images in all directions, remove site/scanner bias, and improve patients' classification and brain developmental volume trajectories in both large and small brain regions. It could be used in further studies that will use the DEFIDIAG data and lead to potential improvement in ID patient care.

Acknowledgments

I would like to personally thank my PhD directors for their advice during these three years and for entrusting me with this project. A special thanks to the members of the jury for accepting our invitation. I'm looking forward to our exchange in November.

Contents

1	Introduction	1
1.1	DEFIDIAG	1
1.2	MR harmonization	2
1.3	Artificial Intelligence - Deep Learning	2
1.4	Strategy adopted	3
1.5	Side projects related to the thesis	5
2	A bibliographic review	7
2.1	Analyzing developing brains	8
2.1.1	Brain development through childhood	8
2.1.2	Autism Spectrum Disorder and brain evolution	9
2.1.3	Co-registration of developing brains	10
2.2	MR image harmonization	11
2.2.1	Statistical methods	12
2.2.2	Generative Deep-Learning harmonization	24
2.2.3	Unlearning Deep-Learning harmonization	28
2.2.4	Methods summary	30
3	ComBat versus cycleGAN for two sites MR images harmonization	33
3.1	Introduction	35
3.2	Materials and methods	35
3.2.1	Data	35
3.2.2	The Combined Association Test: ComBat	35
3.2.3	cycleGAN	36
3.2.4	Experiments	38
3.3	Results	41
3.4	Discussion	45
3.5	Conclusion	46
4	ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization	51
4.1	Introduction	53
4.2	Materials and methods	54
4.2.1	Data	54
4.2.2	ImUnity's model	55
4.2.3	Training	57
4.2.4	Experiments	58

4.3	Results	60
4.4	Discussion	64
4.5	Conclusion	68
4.6	Compliance with ethical standards	68
5	Harmonization impact on brain structure volume and thickness evolution with age on ASD and control subjects	69
5.1	Introduction	71
5.2	Materials and methods	71
5.2.1	Data	71
5.2.2	ImUnity	72
5.2.3	ComBAT	72
5.2.4	Data preprocessing	72
5.2.5	Experiments	73
5.3	Results	77
5.3.1	Volumetric analysis on healthy subjects	77
5.3.2	Thickness analysis on healthy subjects	81
5.3.3	Impact on ASD patients	83
5.3.4	Mixed Linear Model analysis	84
5.3.5	Freesurfer group analysis	86
5.4	Discussion	87
6	Conclusion and perspectives	91
7	Résumé du manuscrit	95
7.1	Introduction	98
7.2	Résumé de chapitre : Bibliographie	100
7.3	Résumé de chapitre : Évaluation de ComBAT et cycleGAN pour l'harmonisation de 2 sites	102
7.4	Résumé de chapitre : ImUnity, un VAE-GAN adapté à l'harmonisation de données IRM multi-centrique	105
7.5	Résumé de chapitre : Proposition d'évaluations des effets de l'harmonisation sur le développement cérébrale chez l'enfant	107
7.6	Conclusion et perspectives	110
8	Personal information	113
8.1	My resume	113
8.2	Scientific production	115
9	Supplementary Material	117
9.1	JFR-2019 Data challenge with Pixyl	118
9.2	DI-Generic: A deep-learning framework dedicated to medical imaging analyse	127
9.3	Colaboration with Kévin Yaou	129
9.4	Teaching & supervising experience	142

List of Figures

2.1	GM & WM volumes evolution	9
2.2	Lenroot brain volumetry results	10
2.3	BHR study - brain volumes evolution	11
2.4	Lange et al. (2015), ASD impacts on males brain volume	12
2.5	Zielinski et al. (2014), ASD impacts on males brain thickness	13
2.6	Retico et al. (2016)	14
2.7	Sanchez et al. (2012a) templates	14
2.8	Harmonization citations	15
2.9	Decile normalization	16
2.10	Shinohara - White-Stripe normalization	18
2.11	RAVEL results	19
2.12	ComBAT results	20
2.13	M&B-ComBat results	22
2.14	Longitudinal-ComBat	24
2.15	DeepHarmony results	26
2.16	CycleGAN	26
2.17	Unlearning module results	29
3.1	cycleGAN for MRI harmonization	36
3.2	Synthetic Global Noises	40
3.3	Synthetic Noise harmonization workflow	40
3.4	Results visualization - Method's effects on global noises	42
3.5	Results visualization - CycleGAN & ComBAT impact	43
3.6	ComBAT and cycleGAN impact on intensity distribution	44
3.7	Harmonization impact - 2 Sites SVM classification	48
3.8	Harmonization impact - 2 Sites SVM classification - AUC	49
4.1	ImUnity's architecture	55
4.2	Gamma contrast modification	59
4.3	Harmonization results on traveling subjects	61
4.4	2.5d median fusion impact - visualization	62
4.5	ImUnity results for multi-scanner harmonization - visualization	64
4.6	Pre-processing impact	65
4.7	ImUnity results on sites and patients classification	66
4.8	Biological module impact	67
5.1	ABIDE subjects age distribution	73
5.2	Total Gray Matter volume evolution	74

5.3	Harmonization metric workflow	75
5.4	ImUnity impact on brain volumes evolution with age for controls	77
5.5	Volume harmonization impact on healthy female - visualization	78
5.6	Thickness and volume harmonization impacts for healthy males - visualization	80
5.7	Harmonization impact on brain thickness evolution with age for controls . .	82
5.8	Thickness and volume harmonization impacts for ASD males - visualization .	84
5.9	Harmonization impact on Freesurfer group analysis on site affiliation	87

List of Tables

2.1	Harmonization methods summary	31
3.1	Harmonization impact - Synthetic global noises classification	42
3.2	Harmonization impact - Synthetic lesions classification	43
3.3	Harmonization impact - Welch's t-test	44
3.4	Results - Pearson's test between radiomic features and site affiliation	44
4.1	Versatility of deep-learning harmonization models	53
4.2	SSIM scores	62
4.3	ImUnity impact on MAE and MSE for multi-scanner harmonization	63
5.1	Harmonization volume metric ratio for healthy subjects	79
5.2	Harmonization thickness metric ratio for healthy males	81
5.3	Harmonization thickness & volume metric ratio for ASD males	83
5.4	KR test harmonization results for brain volumes development	85
5.5	KR test harmonization results for brain thickness development	86

Glossary

ComBAT	<i>ComBined Association Test</i> , statistical harmonization method.
CycleGAN	Deep-learning harmonization model.
Expanded Disability Status Scale	The Kurtzke Expanded Disability Status Scale (EDSS) is a method of quantifying disability in multiple sclerosis..
FreeSurfer	Software used for brain MRI analysis.
ImUnity	<i>Image Unity</i> , deep-learning harmonization model implemented at the GIN.
ROBEX	<i>RObust Brain EXtraction</i> , automatic brain extraction tool for T1-weighted MRI.

Acronyms

AD	Alzheimer Disease.
AI	Artificial Intelligence.
ASD	Autism Syndrome Disorder.
AUC	Area Under Curve.
BMI	Body Mass Index.
CAD	Computer-Aided Diagnosis.
CDF	Cumulative Distribution Functions.
CNN	Convolutional Neural Network.
CSF	Cerebrospinal Fluid.
CT	Computed Tomography.
DA	Domain Adaptation.
DL	Deep Learning.
DTI	Diffusion Tensor Imaging.
FA	Fractional Anisotropy.
GAM	Generalized Additive Model.
GAN	Generative Adversarial Network.
GIN	Grenoble Institute of Neurosciences.
GM	Grey Matter.
GMM	Gaussian Mixture Model.
ID	Intellectual Disability.
IQ	Intellectual Quotient.
MAE	Mean Absolute Error.
MCI	Mild Cognitive Impairment.
MD	Mean diffusivity.
MI	Medical Imaging.
ML	Machine Learning.
MR	Magnetic Resonance.
MRI	Magnetic Resonance Imaging.
MS	Multiple Sclerosis.
MSE	Mean Squared Error.

NAWM	Normal-Appearing White Matter.
PCA	Principle Analysis Component.
PET	Positron Emission Tomography.
RAVEL	Removal of Artificial Voxel Effect by Linear regression.
RF	Radio-Frequency.
ROC	Receiver Operating Characteristic.
ROI	Region Of Interest.
SPIN	Statistical Principles of Image Normalization.
SSIM	Structural Similarity Index.
SVA	Surrogate Variable Analysis.
SVM	Support Vector Machine.
t-SNE	t-distributed Stochastic Neighbor Embedding.
UGA	University Grenoble Alpes.
VAE	Variational Auto-Encoder.
WM	White Matter.

Chapter 1

Introduction

1.1 DEFIDIAG

This PhD project emerged as a side project of the French national project called DEFIDIAG (Binguet et al., 2022). It was supported by MIAI, the Grenoble Multidisciplinary Institute in Artificial Intelligence, within the multiomics chair.

DEFIDIAG, for Intellectual Disability Diagnostic, is a pilot project funded by the program "Plan France Médecine Génomique 2025; PFMG 2025 (PFMG 2025)" and focuses on intellectual disability (ID). Among rare diseases, ID is the leading cause of consultation in pediatric genetic centers and it is estimated that 1 to 3% of the population is ID. It is extremely difficult to diagnose because hundreds of genes are involved in the pathology, which makes it challenging to identify the causes of the disease and adapt to patient care. DEFIDIAG aims 'to demonstrate the feasibility of complete genome sequencing, as well as its effectiveness, first-line, in the determination of genes involved in intellectual disability'. In total, it is expected that 1,275 patients as well as their two biological parents will participate in the study which will last 30 months. Professionals hope to achieve a 60% diagnostic yield using the full genome sequencing. This will 'allow more families to know more quickly the causes of the disease, to benefit in a shortened time from appropriate care and perhaps to prevent the occurrence of specific complications.'

Intrinsically, DEFIDIAG focuses on genetic analyses but the study will also gather brain MR imaging data from subjects across 11 sites all around France. Due to our laboratory's expertise on medical image processing, our PhD project initially aimed at combining these radiomic data with the genetic analysis to gain knowledge on these rare diseases and further improve diagnostics results.

As of today, the DEFIDIAG data is still being collected. In consequence, all of the present work has been devoted to developing image processing tools adapted to the specifics of the DEFIDIAG data and paved the way for future analyses. In particular, 2 major concerns have gained our attention: (1) the subjects in the database are mainly children, and the effect of age on brain development from early childhood to late adolescence must be taken into account during the analyses in order to avoid confounding factors. (2) In the DEFIDIAG context, and in general for all studies focusing on rare diseases, the number of patients may be very small and may limit the robustness of the results. Therefore, it is common to try to pool data coming from different acquisition centers. However, the image acquisition for the DEFIDIAG program did not follow a specific protocol, each inclusion site following its own imaging procedures. This can lead to large unwanted variations in the dataset

because of different scanners and possibly different acquisition protocols. Since the number of patients per site is small, these induced site fluctuations may have larger effects than the biological variations of interest. Methods for correcting for these multi-centric effects are called harmonization methods and have become our main point of focus.

1.2 MR harmonization

Harmonization is a recent and exciting topic in the medical images analysis research field. It concerns all scanning modalities (MR, CT, PET, etc...) and is a direct consequence of the qualitative nature of these acquisitions. For example, standard anatomical MR data (T1w or T2w) acquired from the same patient but at different acquisition sites often lead to different MR images. This is due to the qualitative nature of the acquisitions which produces weighted images (such as T1w or T2w) that are sensitive to technical choices (hardware, sequence parameters) as well as scanner specific properties and artifacts. Consequently, pooling images from multi-center MR studies in order to approach a particular clinical or biological question does not guarantee an increase in statistical power because of a parallel increase in non-biological variance. These unwanted variations in image intensities also prevent large dissemination of machine learning tools that are trained on a specific site and may not generalize their model to other image providers ([Liu et al., 2020](#)). According to us, ‘Harmonization’ refers to the process of **removing site or scanner related noises while preserving biological information in scans**. It is thus a crucial pre-processing step in multi-centric clinical analysis. In the last few years, several image harmonization solutions have been proposed (see [section 2.2](#) for a review) with an increasing focus on nonlinear analyses using Artificial Intelligence tools.

A recent study by [Bottani et al. \(2022\)](#) highlighted important intra-hospital variations in a context of a clinical care study. These results revealed important technical variations within hospitals, either caused by the use of different machines, or due to their evolutions over time. In consequence, it would be very interesting to evaluate the capacity of existing harmonization solutions to be included in routine clinical care.

1.3 Artificial Intelligence - Deep Learning

With the increase of computational power, and the availability of large datasets, Machine Learning (ML) tools have become incredibly popular amongst the medical image community. ML consists in developing computational models that learn from sample data known as ‘training data’ and that relies on mathematical optimization theory, a very active research domain. ML algorithms are capable of recognising different input patterns to make their final predictions. Most popular examples of ML uses are image recognition, natural language processing and computer vision. Support Vector Machine (SVM), Random Forest or K-means algorithms are some very popular ML models and are still considered as ‘state of the art’ nowadays. More recently, as a consequence of the explosion of digital data and computational power, we have witnessed the uprising of Deep-Learning (DL), a sub-group of ML algorithms. DL models are known to be very efficient in learning how to extract relevant features. For example, in computer vision, Convolutional Neural Networks (CNN) have become very famous as they can learn during their training phase how to extract various features from input images, like shapes, edges or textures. As a consequence, CNN have

been used to successfully detect and segment lesions in medical images, used as computer-aided diagnosis (CAD) tools, or to generate images with better SNR or corrected artifacts. A particularly interesting application of DL in the context of multi-centric image harmonization is the possibility to learn image style (such as a particular painter’s artistic style) and to apply this learned style to another unrelated image. Drawbacks of DL models involve their lack of explainability as well as their need for large training datasets. As a consequence, several organizations have been established to gather data coming from different centers in order to maximize dataset’ sample size and thus improving final results. It is thus also directly concerned by harmonization issues.

1.4 Strategy adopted

Given our context, this PhD thesis has focused on the development of new DL tools for medical image harmonization. As the DEFIDIAG data was not available, it was chosen to test our solutions on large available open access imaging datasets, with characteristics as close as possible to the DEFIDIAG one. The ABIDE database [Di Martino et al. \(2014\)](#), which is a large multi-centric database focusing on Autism Spectrum Disorder (ASD) among children, was eventually chosen as a reference. Every study performed during this work was done on the ABIDE database. Two additional databases were used for validation purposes as they contain data from traveling subjects (scanned at different sites) and allow for direct quantitative evaluation.

- **BIBLIOGRAPHIC REVIEW.**

The first part of this thesis consisted in gathering literature results related to our project. It began with a literature review focusing on normal brain development during childhood (see [subsection 2.1.1](#)). This was key for further investigations, as they would be done on children data to match with the initial context. In addition to control subjects, we then turned our attention to neurological brain disorders and their impacts on brain development. Once again, because DEFIDIAG’s main concern is the diagnosis of intellectual disability diagnosis. Finally, we looked at the impact of brain development for brain co-registration to a reference template, which is a common pre-processing step in medical imaging analysis. These templates are meant to represent an average brain of the population under investigation. Because of the large brain variations expected during childhood and adolescence, special templates had to be found.

Our second literature review focused on data-harmonization (see [section 2.2](#)). This topic was first introduced in 1998 and many approaches have been proposed since. Statistical solutions based on histogram matching algorithms were the first ones to give promising results and many variations have been proposed. Subsequently, solutions inspired from the genomic world where the ‘batch effect’ should be removed, have been proposed. ComBAT is one of them, and is still considered today as a reference for MR data-harmonization. Finally, influenced by the increasing popularity of deep learning, models originally developed for image segmentation or generation have been adapted to harmonization. Most recent studies are based on DL, using U-Net, cycleGAN or VAE architectures, and have shown encouraging results.

These points are presented in detail in [chapter 2](#) and a comparative table of existing harmonization techniques is proposed at the end of the chapter.

- **STATE OF THE ART METHODS COMPARISON.**

We present in [chapter 3](#), a study that evaluates two harmonization solutions chosen from the literature. The idea was to compare a recent DL model (cycleGAN), to a reference linear solution (ComBAT) to estimate the potential of DL-based harmonization solutions. Another key point was to propose an original workflow to evaluate harmonization results. In our study, we focused on removing the actual site induced information but also on harmonization impact on synthetic noises that we artificially added to the images. We considered two types of noises: (1) ‘Global noises’ that mimic scanner induced noises and that we wanted to suppress after harmonization, and (2) ‘local noises’ that were assimilated to biological local variation (trauma, stroke, tumor ...), and were meant to be preserved after harmonization. Evaluations were based on extracted radiomic features from images before and after harmonization. Then a SVM classifier was used to detect the presence or absence of signal of interest (site induced, synthetic global or local noise).

- **PROPOSITION OF AN ORIGINAL DL HARMONIZATION MODEL.**

Our first study showed that DL techniques can be adapted to data-harmonization, reaching state of the results under specific conditions. However, as found in the literature review, each DL solution also presents specific requirements due to their intrinsic architecture. For example cycleGAN is limited to two site harmonization only, and seems to produce very good results in this condition. However, this is rarely met in practice and is not adapted to large multi-centric databases. Some other unrealistic requirements involve:

- the need for traveling subjects to fit the solution
- the need to fine tune the solution every time data from a new site or scanner is added
- the need to re-harmonize the data for every clinical application. Ideally, one would want to use the harmonization method once as a tool in the pre-processing workflow.

In [chapter 4](#), we propose an original DL harmonization approach that does not rely on these requirements and is therefore adapted to clinical practice. The new solution is called ImUnity and is based on a VAE-GAN architecture. A significant effort was devoted to validate our tool through different experiments, focusing on site removal effect, diagnosis prediction of ASD and similarity improvement using traveling subjects’ data.

- **HARMONIZATION VALIDATION BASED ON A BRAIN DEVELOPMENTAL STUDY.**

In order to push our validation process further, we eventually tested ImUnity on a clinical application related to the DEFIDIAG program. The study is presented in [chapter 5](#). We evaluated the effect of harmonization on brain apparent volume and cortical thickness evolution during childhood. This was done on both healthy subjects and ASD patients of the ABIDE database, combining the 11 acquisition sites. The age trends generated before and after harmonization were compared to those found in the literature in large mono-centric studies. Part of the study was conducted in collaboration with Constance Sohler, a master student doing her ‘end of study’ internship under my supervision. Our work suggests a positive impact of ImUnity, which reduces site effect while improving biological scores. Further statistical analyses still need to

be performed in order to properly conclude but this type of work could also be used as a generalizable way to evaluate harmonization tools and could be adapted to any multi-centric database.

In conclusion, our investigations on data coming from children from a large age range as well as multiple acquisition sites have led to the development of innovative image processing tools. They have laid the foundations for further studies that will use the DEFIDIAG data and lead to potential improvement in ID patient care.

1.5 Side projects related to the thesis

In parallel with my PhD project, I had the chance to participate in several side projects related to medical image processing or genetic data analysis. These experiences with other labs and scientists are summarized at the end of the manuscript.

From September to October 2019, I was involved in the JFR2019 challenge with [Pixyl](#), a French startup from Grenoble developing AI patient care solutions for different pathologies like Multiple Sclerosis (MS). The challenge consisted in predicting MS patient's EDSS score based on a FLAIR acquisition and several clinical features such as age or sex. My work was to develop a deep learning learning model that takes the FLAIR images, the mask of MS lesions previously segmented by Pixyl solutions and the age of the patient as input. The model's prediction was then pooled with other models (random forest, SVM) results to produce a final ensemble prediction. Eventually, we came in first place in this challenge, in front of very competing teams like Nvidia or Icometrix. A paper detailing this experience was published in [Roca et al. \(2020\)](#), and is presented in supplementary materials [section 9.1](#).

A major point during this thesis was to develop a set of AI tools to analyze large databases. In order to make these tools easily accessible to other scientists, I developed an in-house deep-learning framework named 'dl-generic' which is described in more detail in supplementary materials [section 9.2](#). The purpose of this framework was to enable people with different backgrounds (from students in IT to neurologists) to use deep-learning tools for their personal applications. This framework has already been used in the lab by an engineering student (Yunshi Han) during her 6-months internship focusing on Glioma segmentation and transfer learning. Moreover, Loïc Legris, a neuro-radiologist from CHUGA, with no IT background has also been able to implement DL tools for his own research studies on predicting ischemic stroke using our framework.

As a result of DEFIDIAG duality between genetic and imaging data, and the large scope of interests of the MIAI foundation, I regularly exchanged with other scientists from various backgrounds on our problems. In particular, I collaborated with Dr Kevin Yaury, a geneticist also involved in the DEFIDIAG project. He played a crucial role as an expert to help me better understand the DEFIDIAG purposes. Reciprocally, I helped him in his research project and contributed to one of his publications [Yaury et al. \(2022\)](#), which is presented in supplementary materials [section 9.3](#).

Finally, details about different teaching experiences I had during these 3 years can be found in supplementary materials [section 9.4](#). This includes lectures and practical sessions given at Grenoble INP and a practical course on contrastive learning that I co-organized during the 2022 international winter school AI4Health.

Chapter 2

A bibliographic review

CONTENTS

2.1	Analyzing developing brains	8
2.1.1	Brain development through childhood	8
2.1.2	Autism Spectrum Disorder and brain evolution	9
2.1.3	Co-registration of developing brains	10
2.2	MR image harmonization	11
2.2.1	Statistical methods	12
2.2.2	Generative Deep-Learning harmonization	24
2.2.3	Unlearning Deep-Learning harmonization	28
2.2.4	Methods summary	30

2.1 Analyzing developing brains

The DEFIDIAG project, and more generally studies involving brain images from children and adolescents, present specific challenges. During brain development, brain volumes change globally, but also locally in cortical and subcortical regions. In order to be able to process the data correctly and pool multiple data from different age ranges, we started by looking for references on brain development studies and then on appropriate tools for brain image pre-processing.

2.1.1 Brain development through childhood

Many research teams around the world have investigated the developmental trajectories of human brains during childhood. In this section, we only focus on volume and thickness evolution of different cortical and subcortical brain regions with brain aging. We present mono-centric, longitudinal studies (i.e the monitoring of biological parameters on same patients over a period of time) found in the literature. We investigate brain evolution among healthy controls but also patients presenting ASD to have an idea of structural changes induced by intellectual disorders. These results will also be used in [chapter 5](#) as references to evaluate the effect of image harmonization.

[Giedd et al. \(1999\)](#) present main brain area volumetric evolution through childhood and adolescence, focusing on global cortical regions. This longitudinal study was run on structural MRIs acquired from 145 healthy subjects from 5 to 22 years old (y.o.). Subjects were scanned between 1 and 3 times within a 2 years time lap. Results present a clear shift between males and females results for all brain regions. It suggests that brain development is highly impacted by sex and that this covariate should be taken into account when studying brain development. As presented in [Figure 2.1](#), results also suggest an inverted U-shape tendency for GM volumes during this age range while WM brain volumes tend to increase more linearly over the years. These tendencies have been confirmed by more recent longitudinal studies, ([Brain Development Cooperative Group, 2012](#); [Lenroot et al., 2007](#); [Wierenga et al., 2014](#)) which present brain structures volumetric evolution with age in more details, as they investigated more cerebral regions. In particular, [Lenroot et al. \(2007\)](#) present both cortical or subcortical brain regions volume evolution with age using a large longitudinal pediatric neuroimaging study (829 scans from 387 subjects, ages 3 to 27 years). As shown in [Figure 2.2](#), this study also highlights the need to consider sex matching in studies of brain development and propose equations of brain volumes in function of age and sex, suggesting different model types (linear, quadratic or cubic) according to the observed ROI.

Similarly, the [Brain Development Cooperative Group \(2012\)](#) presents brain volumes evolution for 325 healthy subjects from 4 to 18 y.o. It takes into account biological features like age sex and body mass index (BMI) but also social ones like family income and parental education. Their results, presented in [Figure 2.3](#), suggest that evolution trajectories may differ between males and females (predominantly curvilinear in females and linear in males), a global decrease (resp. increase) of GM (resp. WM) volume over the years, a greater age-related variance in lobar structures and small systematic associations of volumes with BMI but not with IQ, family income, or parental education.

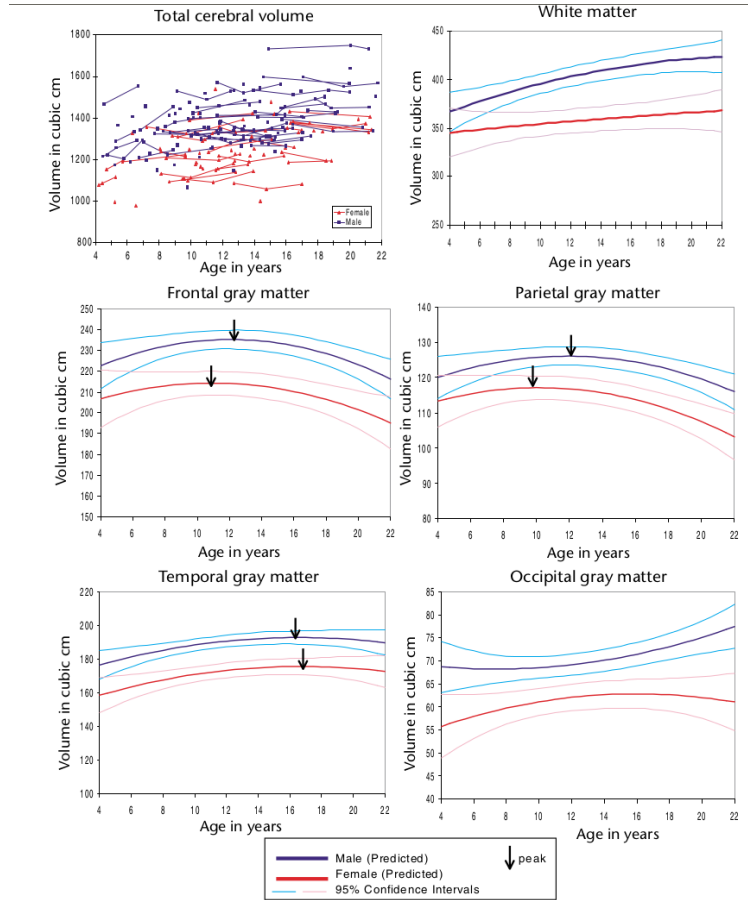


Figure 2.1 – Illustration of main cortical brain area volumetric evolution with age during childhood. Adapted from [Giedd et al. \(1999\)](#); [Lenroot and Giedd \(2006\)](#)

2.1.2 Autism Spectrum Disorder and brain evolution

On top of other covariates, intellectual disorders can also alter structural developmental trajectories in children as reported in [Ma et al. \(2021\)](#). Because the data used in our PhD thesis mainly come from the ABIDE database, we also investigated the impact of ASD on different brain structures evolution. As mentioned in [Langen et al. \(2009\)](#), this impact has long been unclear. For example, [Voelbel et al. \(2006\)](#) reported larger volumes in autism of the caudate nucleus while [Gaffney et al. \(1989\)](#) did not. Other examples reporting opposite results on ASD impact on brain development can be found in the literature. It seems that these differences could be due to the use of neuroleptic drugs that can be associated with volume increases ([Langen et al., 2009](#)). [Lange et al. \(2015\)](#) present a longitudinal study on 100 male participants with ASD and 56 typically developing controls with scans obtained with a 2.7 years interval over an 8-year period. As reported in [Figure 2.4](#), authors suggest a significant reduction of cortical WM mean volume among autistic patients, relatively to the controls. On the contrary, they report an increase of mean ventricular volume across the broad age range studied (6–35 y.o). Their results also suggest that ASD is a dynamic disorder with complex changes over time in whole and regional brain volumes, from childhood to adulthood.

Similarly, [Zielinski et al. \(2014\)](#) present changes in cortical thickness in a longitudinal study gathering 97 male individuals with ASD and 60 typically developing male control sub-

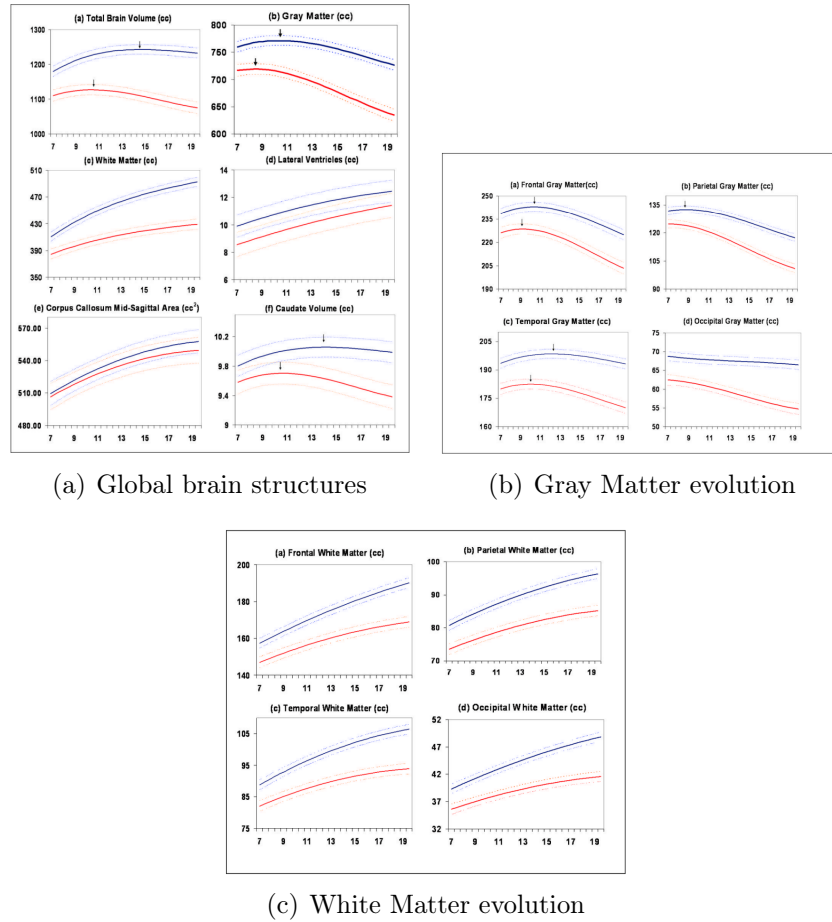


Figure 2.2 – Cortical and subcortical brain structures volumetric evolution with age during childhood and adolescence. Adapted from [Lenroot et al. \(2007\)](#)

jects aged from 3 to 39 y.o. They show significant differences in brain structure thickness for most ROIs. As presented in [Figure 2.5](#), they also show that the impact of ASD varies across the lifespan, affecting ROIs thickness differently in childhood, adolescence, and adulthood.

Due to a significant majority of male subjects in ASD studies, scientists have observed a 3:1 ratio between males and female patients [Loomes et al. \(2017\)](#), which matches the ratio in the ABIDE database. Consequently, published literature studies about females presenting ASD are rarer than for men. [Walsh et al. \(2021\)](#) still reports several studies results on this topic. Similarly to male, ASD seems to affect brain development among females, however the changes seem to be different. For example [Retico et al. \(2016\)](#) report in [Figure 2.6](#) an increase of GM volumes among ASD patients aged from 2 to 7 y.o for both genders but their study also reveals a greater spatial extent in ASD females than ASD males.

2.1.3 Co-registration of developing brains

One crucial point to analyze group images is the 'registration' step, consisting in the alignment of all patient scans to a specific brain of reference also known as 'template'. This template is usually meant to represent a healthy average human brain and the most popular template is the MNI which averages healthy adult brains in specific spatial representations. However, a couple of studies ([Sanchez et al., 2012a,b](#)) have reported that using one of these

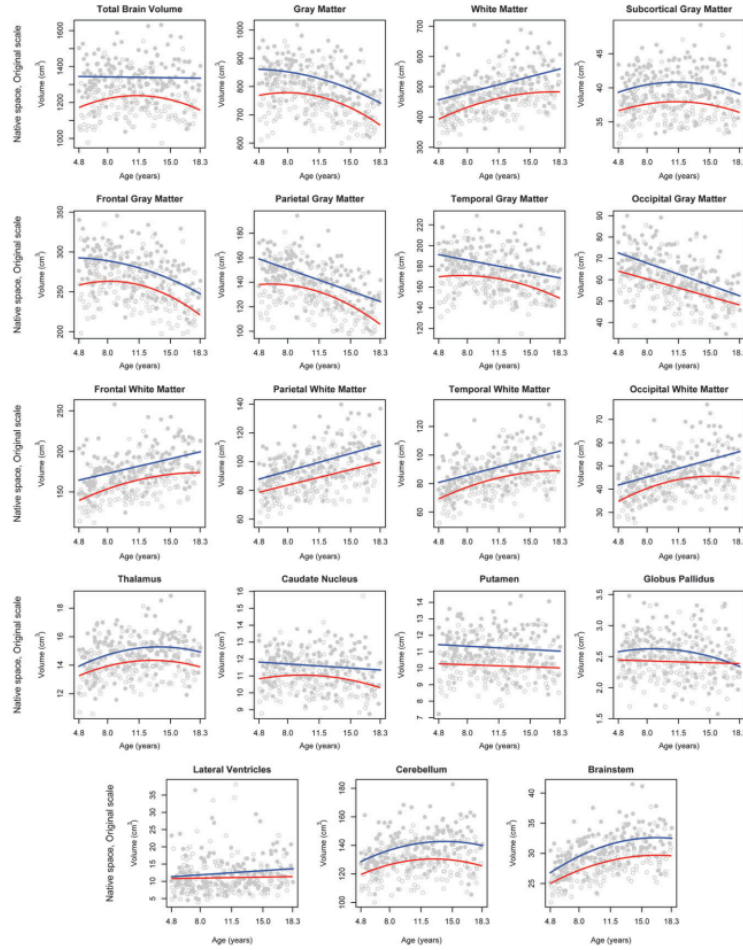


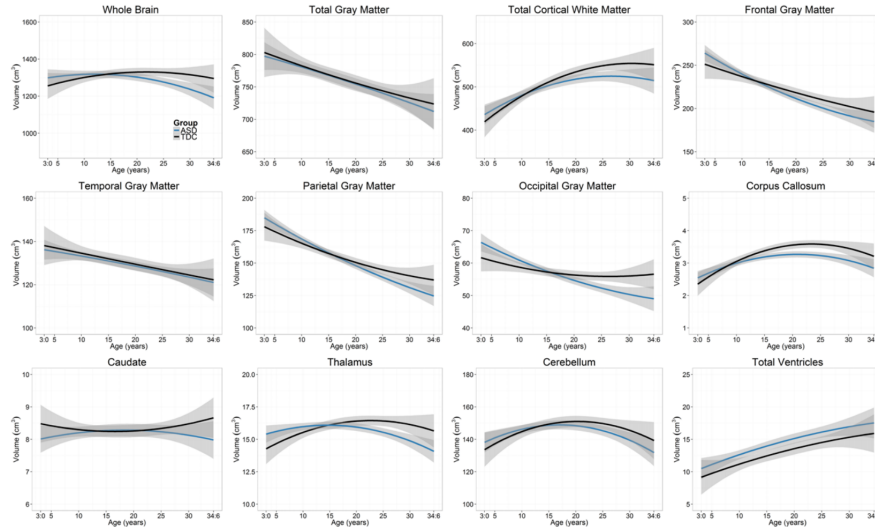
Figure 2.3 – Illustration of brain regions volumetric evolution with age. Adapted from [Brain Development Cooperative Group \(2012\)](#)

templates as a reference for children studies can lead to a loss of information. This is a direct consequence of brain morphology and structural variations with age [subsection 2.1.1](#). Other works ([Wilke et al., 2002, 2003](#); [Yoon et al., 2009](#)) have also pointed out that differences in shape, size, composition and contrast, impact tissue segmentation or diagnosis prediction when co-registering pediatric brains to an adult template. Thus having adapted children's brain templates seems primordial for our study. To tackle this issue, [Sanchez et al. \(2012a\)](#) proposed an open access database containing brain templates in 6 months intervals for the age range from 4.5 to 19.5 y.o. These templates are represented in the MNI-152 space. [Figure 2.7](#) illustrates variations between age specific templates over the years and the pipeline used to create them.

In this PhD thesis, for every experiment involving children's data, we used this open access database of age-specific templates for brain co-registration.

2.2 MR image harmonization

As already mentioned in [section 1.2](#), even after appropriate coregistration of brain images, unwanted variations in image intensity can still corrupt multi-sites analyses. Therefore, MR



(a) Abnormal age-related cortical volume trajectories in ASD.

Longitudinal age and group differences (%)										
Structure	Group (ASD-TDC)	P value	Age	P value	Age × group	P value	Age ²	P value	Age ² × group	P value
Whole brain	n.s.	–	2.25	<0.001	–1.66	<0.001	–0.06	<0.001	0.04	0.005
Total GM	n.s.	–	–0.61	<0.001	n.s.	–	n.s.	–	n.s.	–
Frontal GM	n.s.	–	–0.98	<0.001	n.s.	–	n.s.	–	n.s.	–
Temporal GM	n.s.	–	–0.55	<0.001	n.s.	–	–0.03	<0.05	0.02	<0.05
Parietal GM	n.s.	–	–0.32	<0.021	n.s.	–	n.s.	–	–0.01	<0.05
Occipital GM	n.s.	–	–0.44	0.002	–0.15	<0.01	n.s.	–	0.02	0.002
Total cortical WM	–3.91	0.005	3.90	<0.001	–0.53	0.002	–0.07	<0.001	n.s.	–
Corpus callosum	n.s.	–	3.42	<0.001	n.s.	–	–0.07	<0.001	n.s.	–
Caudate	n.s.	–	1.15	<0.001	n.s.	–	–0.03	<0.001	n.s.	–
Thalamus	n.s.	–	3.15	<0.001	n.s.	–	–0.08	0.001	n.s.	–
Total cerebellum	n.s.	–	3.25	<0.001	n.s.	–	–0.06	<0.001	n.s.	–
Total ventricles	n.s.	–	1.33	<0.001	n.s.	–	n.s.	–	–0.01	<0.05

GM, gray matter; n.s., not significant at $\alpha = 0.05$.

(b) Percent changes in group mean growth curves between controls and autistic subjects

Figure 2.4 – Impacts of ASD on brain ROIs volume throughout lifespan. Adapted from Lange et al. (2015).

image harmonization will play a crucial part in the DEFIDIAG study. As shown in Figure 2.8, the interest for MR harmonization has increased over the last 30 years, probably due to an increase of multi-centric studies. In the next sections, we will point out the main challenges of harmonization, and how the field has evolved from classical statistical solutions to more complex solutions with the popularization of DL approaches. Additionally, we will summarize current technical solutions and understand their main strengths and weaknesses.

2.2.1 Statistical methods

As presented in Sled et al. (1998), the first papers that have looked at the issue of scanner-induced bias in multi-site MR studies have focused their attention in reducing extreme variations seen in surface coil images. With the growing interest in automatic segmentation, new methods involving 'histogram matching' or 'intensity non-uniformity correction' algorithms have been developed. As a result of these studies, the N3 algorithm (a non-parametric approach for field bias intensity correction) was presented. Later on, more complex solutions have been proposed, each time trying to tackle previous solution's drawbacks. Most statis-

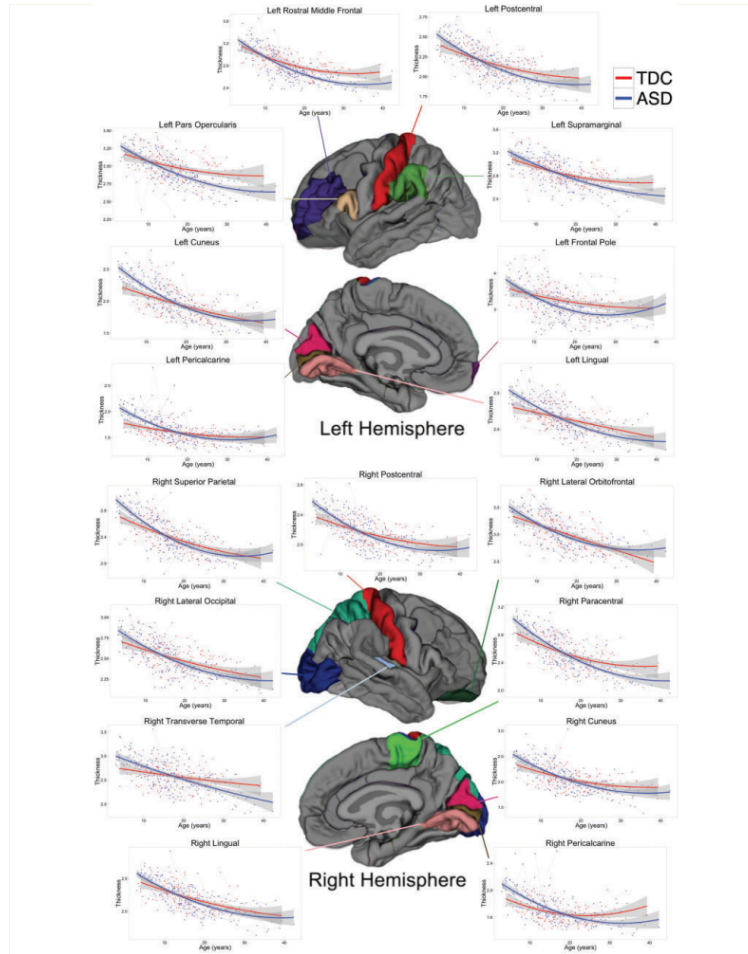


Figure 2.5 – Impacts of ASD on brain ROIs thickness throughout lifespan. **Top:** coloured brain regions identify significant group differences in age-related cortical thickness changes. **Bottom:** Regional differences in group mean cortical thickness during childhood, adolescence, and adulthood. Values are percentage differences in ASD relative to typically developing control subjects. R = right; L = left. Adapted from [Zielinski et al. \(2014\)](#)

tical solutions published in the literature focus on image intensity distribution and usually involve histogram matching algorithms and its revised versions.

Principle of histogram matching

Histogram matching is a classical processing step in image analysis and is used to adjust the contrast of one image according to the contrast of another image. It is based on both images' intensities distribution.

This principle consists in a few steps:

1. Extract brain tissue voxels of interest, as the intensities of background or cranial voxels are not meant to be aligned.
2. Considering every selected voxel, compute histograms for the reference image and the

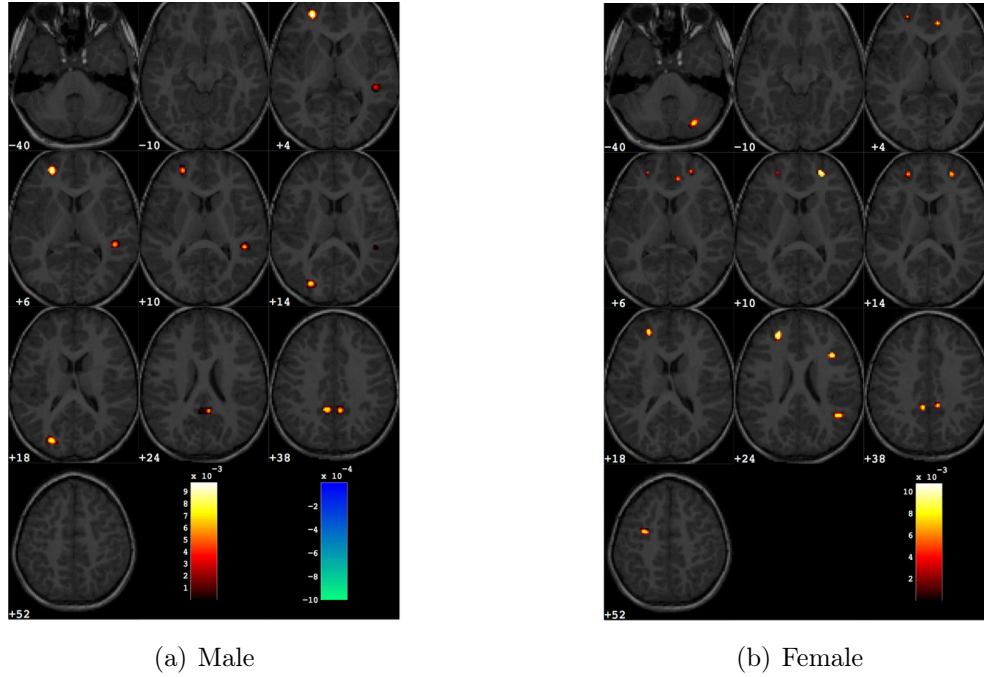


Figure 2.6 – Discrimination map highlighting the differences between subjects with ASD from matched controls. The regions in red scale represent the brain areas where GM is greater in groups with ASD with respect to controls. Figures adapted from [Retico et al. \(2016\)](#)

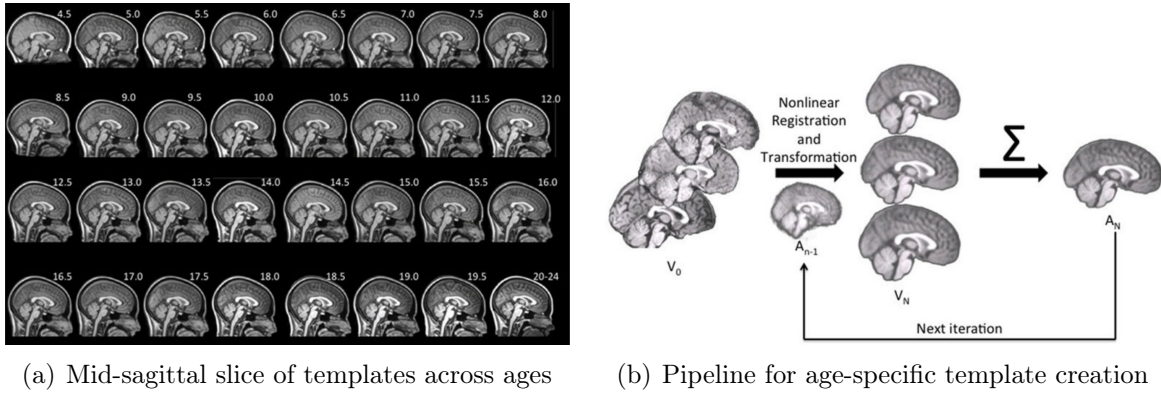


Figure 2.7 – Brain Templates across age. a) Templates visualization and b) pipeline representation used for their creation. Figures adapted from [Sanchez et al. \(2012a\)](#)

ones to harmonize as follows:

$$h(i) = \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \sum_{z=1}^{N_z} \begin{cases} 1 & \text{if } f(x, y, z) = i \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

3. Compute histograms cumulative distribution functions (CDF):

$$CDF(j) = \sum_{i=1}^j h(i) \quad (2.2)$$

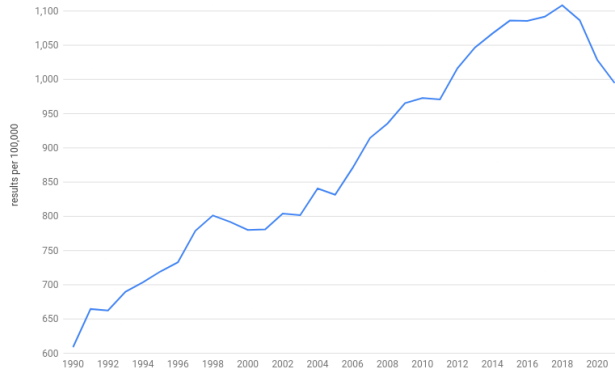


Figure 2.8 – Results per 100,000 citations in PubMed for the query: (*'mri' OR 'ct' OR 'mr'*) AND (*'harmonization' OR 'normalization'*)

4. Once the CDF functions for the image and reference are known, for each gray level G_1 (usually $G_1 \in [0; 255]$) find G_2 such that $CDF_{img}(G_1) = CDF_{ref}(G_2)$. This gives the final histogram matching function $M(G_1) = G_2$ that is applied to all voxels values.

To our knowledge, [Wang et al. \(1998\)](#) published the first study that applied histogram matching to MR harmonization. Using a version of this algorithm that considered histogram bins, they could improve MS lesion volume measurements.

Histogram matching based solutions

Several years later, [Shah et al. \(2011\)](#) presented a succinct review of existing published MR harmonization solutions, pointing out some crucial points faced when developing harmonization solutions. The current statistical methods were derived from intensity histogram matching algorithms as presented above in [section 2.2.1](#) and included even-order histogram matching ([Christensen, 2003](#)); multiplicative correction field for matching a template histogram to a reference model density by minimizing KL divergence between distributions ([Weisenfeld and Warfteld, 2004](#)); region specific normalization using GMM ([Hellier, 2003](#)); and an intensity normalization method mapping the problem to an image registration one ([Jager et al., 2006](#)). Even if these solutions had first shown promising results in their specific context, [Shah et al. \(2011\)](#) point out several points limiting their use and generalization. First, they were not validated to unseen domains, therefore limiting their use to the ones used to fit the models. Similarly, validation on various pathologies were missing. This stressed out for the first time the real challenge when developing harmonization solutions: the validation part is not straightforward and can be complex. Additionally, some solutions were found to be too slow, clearly limiting their use as a pre-processing step when working with large databases. Finally, a challenge for harmonization solutions was to improve ‘any’ multi-center clinical analyse, meaning that the technique had to generalize its training to all unseen data and had to be adapted to any pathology that can locally impact images intensity (lesions might induce local hyper or hypo intensities variations for example).

Subsequently, [Shah et al. \(2011\)](#) presented a new solution with better performances in terms of speed, usability and area generalization. The ‘Decile normalization’ algorithm was first introduced in [Nyul et al. \(2000\)](#), and had already been chosen in various clinical applications as an intensity normalization step. It consists in an histogram matching solution discretized by percentiles followed by a control that all harmonized images have the same

percentiles values. This technique is less strict than classical histogram matching algorithms, and allows better inter subject local variations which have to be preserved as they reflect natural variation between all subjects. It was shown that ‘Decile normalization’ (Nyul et al., 2000) could reduce inter-scanner variability, resulting in more homogeneous intensity values for voxels of the same tissue type. Harmonization could also improve MS lesion segmentation, improving the Dice score of every segmentation method, with a greater impact of the decile method when more complex classifiers are used (Bayesian classifier). Same results were found for tissue types classification, meaning a better differentiation of the lesions but also of the tissues after decile harmonization. Figure 2.9 illustrates the workflow of the ‘Decile normalization’ method and the results obtained by Shah et al. (2011) using it.

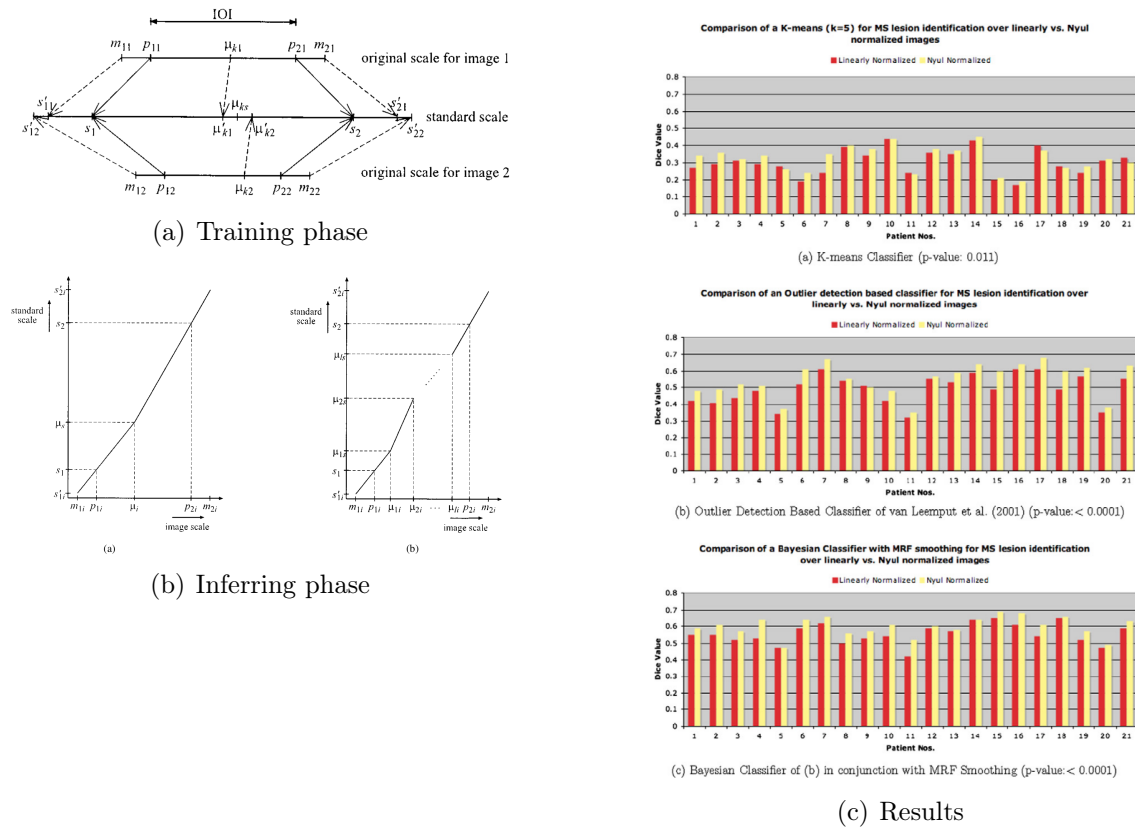


Figure 2.9 – Decile normalization. **A)**: The figure illustrates the training stage with the intensity landmarks from 2 input images (top and bottom) mapped to the standard scale (middle). The standard scale intensity landmarks are then obtained by taking the means of the mapped intensity landmarks. **B)**: The figure illustrates the transformation stage where a new input image histogram (on the horizontal axis) is mapped to the standard scale in a piecewise linear fashion. **C)**: Results comparing the effects of linear normalization (red bars) to the decile normalization (yellow bars) on various MS lesion identification approaches. Dice scores were computed with a consensus labeling of 5 experts. Illustrations adapted from (Nyul et al., 2000; Shah et al., 2011).

The legacy of histogram matching

Later on, and in view of the last results, [Shinohara et al. \(2014\)](#) listed for the first time 7 principles a harmonization solution should meet. These principles form the 'Statistical Principles of Image Normalization' (SPIN) and are listed below. A good harmonization method should:

1. have a common interpretation across locations within the same tissue type;
2. be replicable;
3. preserve the rank of intensities;
4. have similar distributions for the same tissues of interest within and across patients;
5. not be influenced by biological abnormality or population heterogeneity;
6. be minimally sensitive to noise and artifacts;
7. not result in a loss of information associated with pathology or other phenomena;

According to the authors, previously mentioned histogram matching based solutions 'suffer from the lack of biological interpretability of the normalized units'. It also seems that 'Decile Normalization' ([Nyul et al., 2000](#); [Shah et al., 2011](#)) violates a couple of SPIN principles. For example, due to some strong assumptions (the distribution of tissue-type is the same across subjects and visits; subjects' brains do not have abnormal pathology; technical artifacts do not exist), variations in intensities are difficult to interpret, making any histogram-matching method inappropriate for any study of images from multiple subjects. In their study, [Shinohara et al. \(2014\)](#) report a 'false erosion of GM on a magnitude much larger than would be expected' (in the case of Alzheimer's disease (AD) patients for example), meaning that such normalization would lose this relevant variation.

White-Stripe normalization

In 2014, [Shinohara et al. \(2014\)](#) presented a novel normalization solution in accordance with the SPIN principles. It is an adapted tissue-specific histogram matching normalization. More precisely, they presented the "White-Stripe normalization" protocol able to match moments of the white matter (WM) voxels intensity distribution. The method applies a z-score transformation to all voxels intensity using parameters estimated from Normal-Appearing White Matter (NAWM) intensities. Unlike classical histogram matching methods, it satisfies SPIN as it is designed to be robust to artifacts and pathologies which is a crucial point when working with multi-centric data presenting lesions (MS, glioma, etc...). They even proposed a hybrid extension algorithm when multi-modality images are available, normalizing images using tissue from the white stripe in all modalities. They showed great improvements compared to classical histogram matching approaches, especially for GM regions on raw image that disappeared after histogram matching but are preserved by their algorithm as shown in [Figure 2.10](#). Authors validated this solution on different datasets involving AD, MS and healthy subjects from several multi-centric databases. The proposed hybrid solution could reduce intra-tissue variance while preserving inter-tissue variance without worsening lesion detection. This technique remains widely used in research as a pre-processing step and will also be used in further experiments presented in this thesis [section 4.2](#).

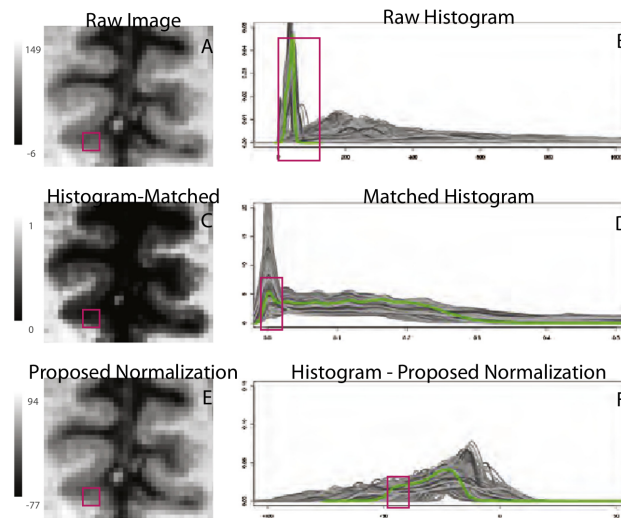


Figure 2.10 – White-Stripe normalization. Impact of different normalization methods showing the failure of histogram matching methods in preserving GM regions. The red square indicates the region of gray matter on the raw image that disappears after histogram matching. The green line shows the histogram for the image shown in the left column. Illustrations taken from (Shinohara et al., 2014).

The RAVEL method

Even if the White-Stripe normalization method showed great results, Fortin et al. (2016) and Dewey et al. (2019) reported that it does not remove all unwanted variations impacting gray matter voxels intensity distribution. It is in fact an overall inconvenience for all histogram matching based methods which have the unfortunate tendency to skip local contrast information and instead assume global histogram correspondence between images. This can be problematic in cases including subjects with pathological differences. Global biological variations can affect the proportions in the global histogram, without any change in contrast. This is why Fortin et al. (2016) proposed an extension of 'White-Stripe' normalization previously described called 'Removal of Artificial Voxel Effect by Linear regression' (RAVEL). For the authors, the normalization step was necessary but not sufficient for a fully harmonization procedure. The RAVEL technique is inspired by three genomics studies (Gagnon-Bartsch and Speed, 2012; Leek and Storey, 2007, 2008) and aims at removing remaining scanner bias after intensity normalization. To do so, they estimate the remaining unwanted variation in control regions and then correct this effect on all brain regions. They choose the cerebrospinal fluid (CSF) as reference as it is highly unlikely that its intensities would be associated with any disease status Fortin et al. (2016). They validated this solution showing greater correlation between voxels intensity and AD status after RAVEL. This suggested a positive impact of RAVEL correction on the discovery of brain regions associated with disease and could facilitate the development of biomarkers using MRI intensities. In fact, as shown in Figure 2.11, the RAVEL method was able to align intensity distributions better than previously mentioned methods and could enhance the classification metrics of AD patients. However, as it strongly suggests that the chosen control region does not carry any biological information, it may be inappropriate when there are great demographic differences between population inter-sites. For example, when groups have different aging distributions, then RAVEL will remove this biological signal of interest, as age affects CSF voxels intensity. An even worse

scenario would be if the control region is directly associated with the outcome of interest of the study. On the other hand, it is also primordial that the control region carries information about sites or scanners otherwise it would not be able to remove it properly in other brain regions.

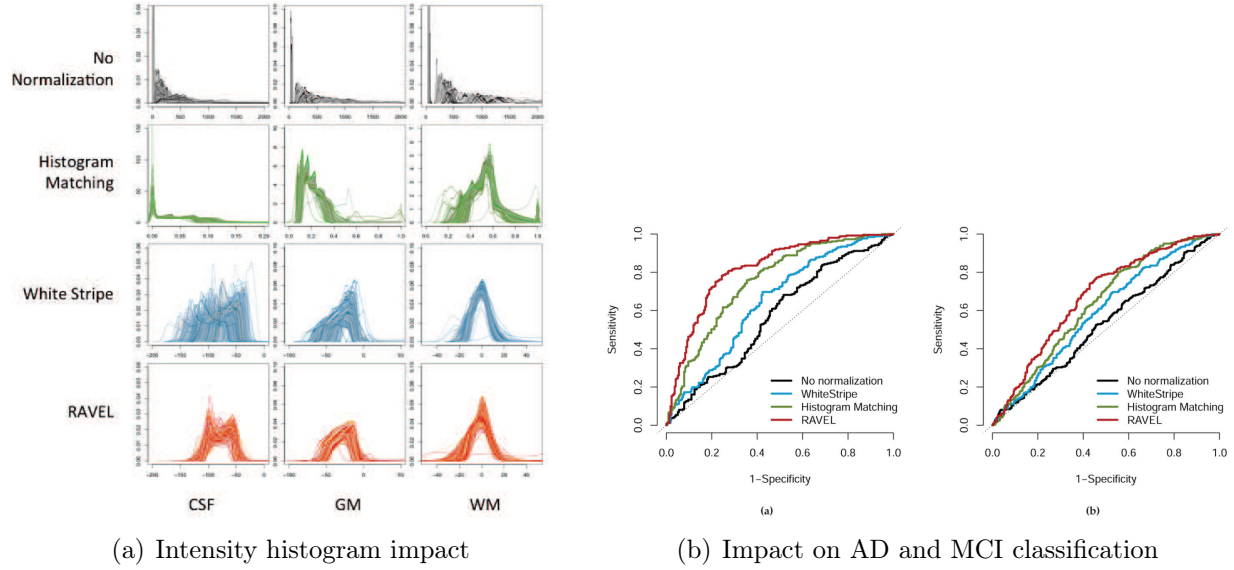


Figure 2.11 – RAVEL. **A)**: Different pre-processing steps effect on CSF, GM and WM intensity histogram. **B)**: RAVEL improves ROC curves for Alzheimer’s disease (AD) and Mild Cognitive Impairment (MCI) classifiers. Predictions were based on mean hippocampus voxel intensity, considering a threshold to classify subjects’ status. Illustrations adapted from (Fortin et al., 2016).

ComBAT: the gold standard?

The Combined Association Test (ComBAT) was first proposed by Johnson et al. (2007) to reduce batch effects on genetic data. It was then adapted for diffusion imaging harmonization by Fortin et al. (2017). The ComBAT model can be summarized as follows. Presuming that data come from m imaging sites, with n_i scans ($i = 1, 2, \dots, m$), for every voxel position v of scan j acquired at site i , the intensity y_{ijv} is modeled as below :

$$y_{ijv} = \alpha_v + X_{ij}\beta_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv} \quad (2.3)$$

Where α_v is the overall intensity measurement for voxel v , X is the matrix of biological co-variates of interest (age and sex in further chapters) and β_v a vector of regression coefficients corresponding to X at voxel v . The model assumes that the error term ϵ_{ijv} follows a normal distribution $\mathcal{N}(0, \sigma^2)$. γ_{iv} and δ_{iv} represent unwanted terms to be removed, and follow normal $\mathcal{N}(\gamma_i, \tau_i^2)$ and Inverse-Gamma(λ_i, θ_i) distributions respectively. Model parameters are updated through empirical Bayes iterations to reduce their variance. Finally, a statistical distribution is obtained for each parameter, allowing to remove the unwanted information:

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \hat{\alpha}_v - X_{ij}\hat{\beta}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + \hat{\alpha}_v + X_{ij}\hat{\beta}_v \quad (2.4)$$

In a subsequent study, the team that proposed RAVEL published a review comparing several existing solutions for Diffusion Tensor Imaging (DTI) derived metrics (FA and MD) harmonization (Fortin et al., 2017). In this review, the authors highlighted the need for harmonization and presented 5 harmonization solutions: ‘Global Scaling’, ‘Functional normalization’, ‘RAVEL’, ‘SVA’ and ‘ComBAT’. They compared them through different validation experiments, and reported that ComBAT performed best at modeling and removing the unwanted inter-site variability in FA and MD maps. In fact, as presented in Figure 2.12, their study showed that ComBAT both preserved biological variability and removed the unwanted variation introduced by site much better than other techniques. It was the first harmonization solution designed to model both biological and unwanted variations.

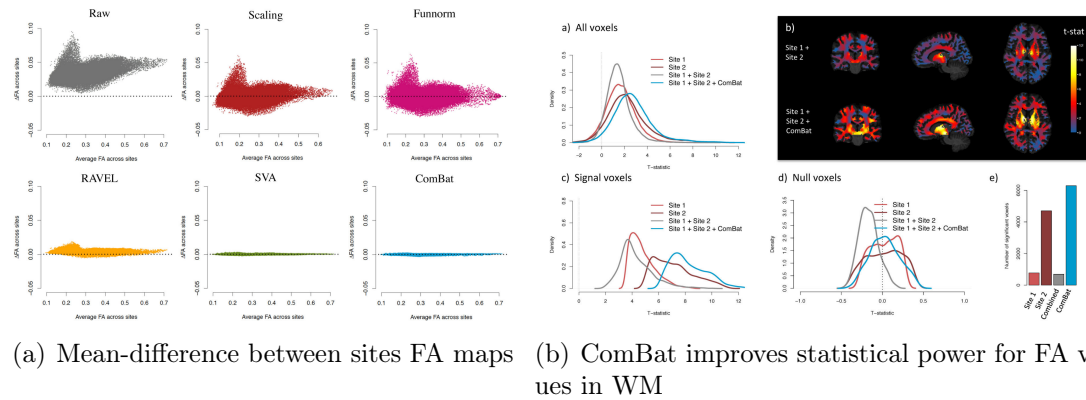


Figure 2.12 – ComBat. **A)**: Mean-difference plot for the FA maps for the different harmonization methods. At each voxel in the WM, the y-axis represents the difference between the average FA value at site 1 and the average FA value at site 2, and the x-axis represents the average FA value across all participants from both sites. **B)**: Voxel-wise t-statistics in the WM, testing for association between FA values and age. 4 combinations of the data: Dataset 1 and Dataset 2 analyzed separately, Dataset 1 and Dataset 2 combined without any harmonization, and Dataset 1 and Dataset 2 combined and harmonized with ComBat. Illustrations adapted from (Fortin et al., 2017).

Although it relies on a strong hypothesis for parameters prior distributions, ComBAT is known to be robust to small sample sizes and is considered as the state of the art statistical technique for diffusion images harmonization. Furthermore, this solution has been widely used for MR harmonization in general (Acquitter et al., 2022; Bell et al., 2022; Eshaghzadeh Torbati et al., 2021; Mahon et al., 2020; Orlhac et al., 2019).

Note that ComBat can be used directly on image intensities to generate harmonized images but is also very efficient in harmonizing derived metrics like volumes, thickness or radiomic features. Several studies report great effect on multiple MR sequences (Bell et al., 2022; Eshaghzadeh Torbati et al., 2021), but also on MR derived metrics like radiomic features (Acquitter et al., 2022; Mahon et al., 2020; Orlhac et al., 2019). More details on how to use ComBAT for direct T1w images harmonization are presented in chapter 3.

Also note that authors developed an user-friendly open-access python module ‘**NeuroCombat**’ allowing users to use ComBAT on their own datasets. Even if its use requires to be fitted for every new site or scanner, ComBAT is still considered as a reference in terms of harmonization, mostly for being easy and fast to use, and quite efficient on small sample size datasets.

ComBAT based solutions

After the original publication, ComBAT has been widely used, and other derived versions of the code have been proposed [Beer et al. \(2020\)](#); [Da-ano et al. \(2020\)](#); [Pomponio et al. \(2020\)](#):

- **COMBAT-GAM**

In their study, [Pomponio et al. \(2020\)](#) proposed a non-linear ComBat version based on

the Generalized Additive Model (GAM). This was motivated by the fact that age has a non-linear effect on brain development (see [subsection 2.1.1](#)). Instead of modeling biological effects on voxels intensity linearly, as done in [Equation 2.3](#). They substituted for it a GAM which is a function of the covariates (age, sex, and ICV), to allow for nonlinear age trends in ROI volumes informed by the data. Replacing the term $\hat{\alpha}_v - X_{ij}\hat{\beta}_v$ by $f_v(X_{ij}, Z_{ij}, \omega_{ij}) = \hat{\alpha}_v + g(X_{ij}) + b_v * Z_{ij} + c_v * \omega_{ij}$, with g being an estimated smoothed nonlinear function for age. This leads to the final equation for ‘ComBat-GAN corrected’ intensity voxels:

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - f_v(X_{ij}, Z_{ij}, \omega_{ij}) - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} + f_v(X_{ij}, Z_{ij}, \omega_{ij}) \quad (2.5)$$

Gathering 10,477 healthy scans from 18 sites, authors reported a positive impact of this non-linear ComBat version. For age prediction, their method led to better harmonized data than the linear-ComBat, as authors obtained the smallest Mean Absolute Error (MAE) with the ComBat-GAM harmonized data.

- **M-COMBAT**

At the same time, [Da-ano et al. \(2020\)](#) proposed 3 original ComBat versions: ‘M-

ComBat’ and ‘B-ComBat’ and the hybrid ‘BM-ComBat’. The first one was initially developed for genetic batch effects removal and presented by [Stein et al. \(2015\)](#). The main idea relies in changing the standardizing mean and variance of the estimates, $\hat{\alpha}_v$ and $\hat{\delta}_v$ to center-wise estimates, $\hat{\alpha}_{iv}$ and $\hat{\delta}_{iv}$, such that the standardized values are then given by

$$Z_{ijv} = \frac{y_{ijv} - \hat{\alpha}_{iv} - X_{ij}\hat{\beta}_v}{\hat{\delta}_{iv}} \quad (2.6)$$

Then the user chose a site $i=r$ as the reference and the final M-ComBat adjusted voxels intensities are:

$$y_{ijv}^{M-ComBat} = \frac{\hat{\alpha}_{rv}}{\hat{\delta}_{iv}} \left(Z_{ijv} - \hat{\delta}_{iv} \right) + \hat{\alpha}_{rv} + X_{ij}\hat{\beta}_v \quad (2.7)$$

- **M&B-COMBAT**

In their study, [Da-ano et al. \(2020\)](#) also presented the B-ComBat version, ‘B’ standing

here for ‘Bootstrap’. This version is also compatible with the ‘M-ComBat’ presented above. It consists of resampling (B=1000 times) with replacement of the initial estimates obtained in the fitted ComBAT ([Equation 2.4](#)). Once the B estimates for each coefficient are obtained, the final estimates of the coefficients are computed using Monte Carlo method by getting the mean of the B estimates. Then the user simply replaces previous estimates in [Equation 2.4](#) or [Equation 2.7](#) by the ones obtained. In their study,

the authors tested these methods on radiomic features extracted from 2 multi-center clinical studies, gathering in total 217 scans from 8 different sites. In all cases, authors reported a successful site-differences removal on observed radiomic features. The proposed ComBAT version could also slightly but consistently improve the predictive power of resulting radiomic models.

Figure 2.13 presents the main results of this study, highlighting the positive effect of all harmonization methods and the interest of the proposed ComBAT version.

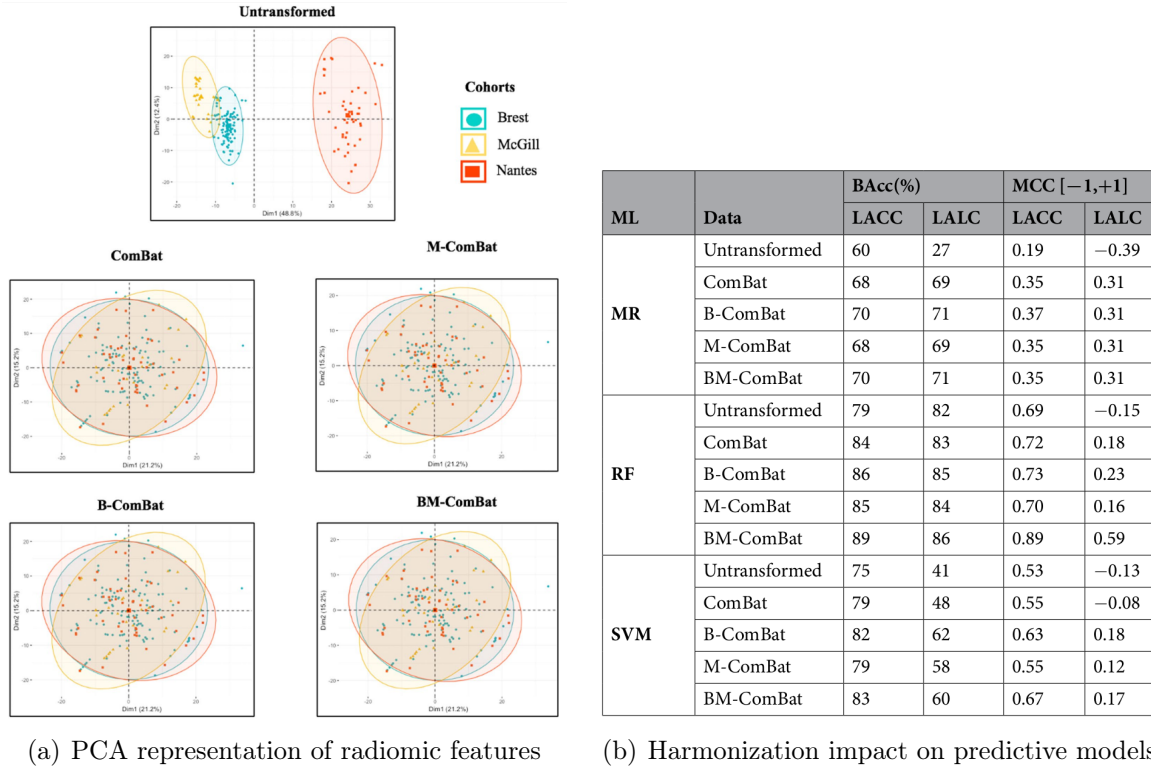


Figure 2.13 – M&B-ComBat. **A)**: Scatter plots of top 2 principal components of the radiomic features across the three labels (centers from the first multi-centric database) using untransformed data or data transformed with the 4 versions of ComBAT. **B)**: Performance metrics evaluation of predictive models in both multi-center datasets (LACC and LALC) testing sets using the three ML pipelines. Presenting the ‘Balanced accuracy’ (BACC) and ‘Matthews correlation coefficient’ (MCC, worst value = -1; best value = +1) metrics for comparison. Illustrations adapted from (Da-ano et al., 2020).

• LONGITUDINAL-COMBAT

The last updated ComBat-version presented in the literature is called ‘Longitudinal

ComBat’. As indicated, it has been developed by Beer et al. (2020) for longitudinal multi-center studies harmonization. As for the ‘M-ComBat’ algorithm, the main modification occurs in the standardization step. To properly account for the dependence of repeated within-subject observations, authors proposed to use a feature-wise linear mixed effects model with a random subject-specific intercept $\zeta_{jv} \sim N(0, \rho_v^2)$. They estimate the fixed effect parameters $\alpha_v, \beta_v, \gamma_{iv}$ using the best linear unbiased estimator (BLUE), the subject random effect variance ρ_v^2 and error variance σ_v^2 with the restricted

maximum likelihood (REML) estimator, and subject-specific intercepts ζ_{jv} using the best linear unbiased predictor (BLUP). We finally have the following standardization algorithm at time t :

$$Z_{ijv}(t) = \frac{y_{ijv}(t) - \hat{\alpha}_{iv}^{BLUE} - X_{ij}(t)\hat{\beta}_v^{BLUE} - \hat{\zeta}_{jv}^{BLUP}}{\hat{\delta}_{iv}^{REML}} \quad (2.8)$$

Giving finally, for the final longitudinal-ComBat correction the equation:

$$y_{ijv}^{long-ComBat}(t) = \frac{\hat{\alpha}_{iv}^{REML}}{\hat{\delta}_{iv}} \left(Z_{ijv}(t) - \hat{\delta}_{iv} \right) + \hat{\alpha}_{iv}^{BLUE} + X_{ij}(t)\hat{\beta}_v^{BLUE} + \hat{\zeta}_{jv}^{BLUP} \quad (2.9)$$

The ADNI database was used for this study. It gathers data from 663 subjects (197 controls, 324 presenting late MCI, and 142 AD patients) across 58 sites involving 152 different scanners. Structural MRI brain scans were acquired at 6 or 12 month intervals for up to 3 years from baseline, and many participants have been scanned on different scanners across visits. Accordingly, authors highlight the crucial need for harmonization when dealing with this multi-centric database. As reported in [Figure 2.14](#) for the frontal cortical thickness, scanners have significant additive and multiplicative effects on the observed features before harmonization. These differences remain after ComBat harmonization but not after longitudinal ComBat harmonized data.

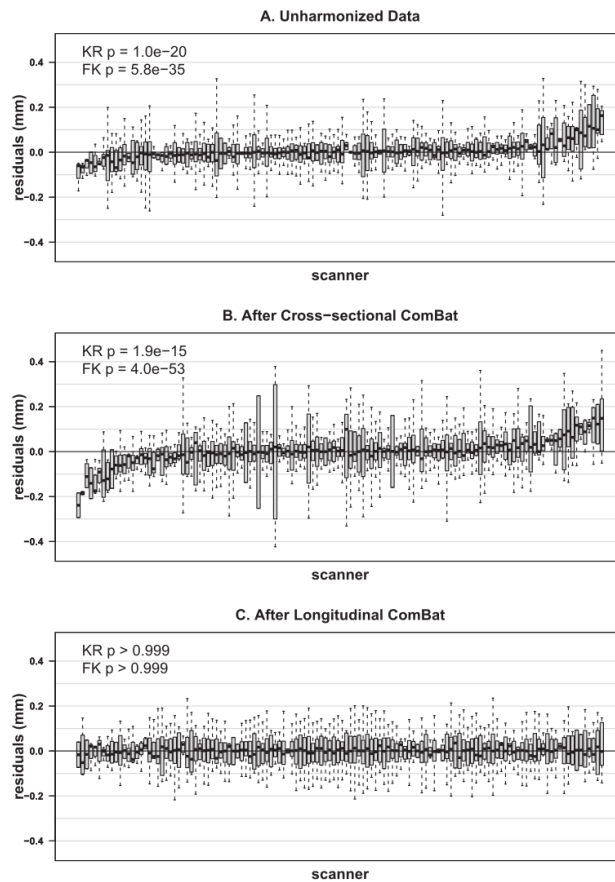


Figure 2.14 – Longitudinal-ComBat. Distributions of left superior frontal cortical thickness residuals across scanners before harmonization (A), after cross-sectional ComBat (B), and after longitudinal-ComBat (C; REML method). Showing in each case, the Kenward-Roger (KR) (resp. Fligner-Killeen (FK)) p-value testing for additive (resp. multiplicative) scanner effects. Illustration taken from (Beer et al., 2020).

2.2.2 Generative Deep-Learning harmonization

Over the last 10 years, deep-learning has been widely used in medical imaging studies in general. It is also true for MR-harmonization which can be seen as a domain adaptation (DA) problem.

In this section, we will focus on generative DL solutions. As their name suggests, these models aim at removing unwanted signals as well as generating ‘corrected’ harmonized images. Generative solutions can be seen as pre-processing steps, and once done, the harmonized scan can be used in any clinical process. To be efficient, these solutions should be adapted to any site or scanner and should not require additional information than the scan to be harmonized. Otherwise their integration to the clinical pipeline would be compromised. With the growth of DL, many new generative techniques have been proposed, and are, for most of them, based on GAN Goodfellow et al. (2014) and U-Net Ronneberger et al. (2015) architectures.

- **DEEPHARMONY** The first DL model adapted to MR-harmonization is called ‘DeepHarmony’ Dewey et al. (2019). It is based on a U-Net Ronneberger et al. (2015) model architecture but the authors applied some variations to the original architecture as proposed by Zhao et al. (2017). DeepHarmony was tested on two databases. The first one

gathers 12 traveling subjects (10 subjects with MS and 2 healthy subjects) that were scanned twice within 30 days on two separate Philips Achieva 3T scanners. The second one is a longitudinal database collected from 45 patients with relapsing remitting MS over 10 years presenting acquisitions with the same two protocols. Each acquisition included a T1-weighted image, a T2-FLAIR image and a dual-echo PD-/T2-weighted image.

The model was trained to match contrast variations induced by the two different acquisition protocols. As traveling subjects were used in this study, ground truth was available. Therefore, the MAE loss could be directly used during training. As mentioned in the paper, they also compared the model’s ability to correct for protocol variations when all contrasts available (4 in the present study) were given as input. Additionally, the authors proposed to use the model in a 2.5D way, which consisted in training 3 equivalent models along the 3 natural axes (axial, sagittal, and coronal). They proposed to combine prediction using the voxel-wise median as it seemed more robust to outliers than the mean.

The authors reported very promising results, as shown in [Figure 2.15](#) where they were able to reduce induced protocol bias and to align brain atrophy trajectories for longitudinal data acquired with both protocols. They also demonstrated the benefit of using a 2.5D approach instead of a classical 2D model. Similarly, it is interesting to note that the use of multiple contrasts as input helped the model to produce better harmonized outputs.

DeepHarmony was the first DL study tackling directly MR-harmonization, and showed the great potential of DL in general for this field of study. However, the fact that its training requires traveling subjects data is not viable for clinical practice. Moreover, this solution has been tested on only two given protocols, and it is very unlikely that its training would generalize to unseen protocols.

- **CYCLEGAN** is a deep-learning model originally developed in 2016 to resolve Image-to-Image translation tasks in [Zhu et al. \(2018\)](#). As illustrated in [Figure 2.16](#), its principle relies on two Generative Adversarial Models (GAN) learning how to map images translation in opposite order (GAN1 : $A \rightarrow B$; GAN2 : $B \rightarrow A$). Here A and B refer to two different sites or scanners. In other words, each GAN will learn how to map an image from one site to its equivalent in the second site domain representation. When the training is successful, it is possible to recover the original input at the output of the second GAN. In the context of data harmonization, an important feature of this model is that its training is unsupervised. In fact, no ground truth (traveling subject) is required for training.

In [chapter 3](#), we evaluate its performances in different contexts and compare it to ComBAT. Refer to [subsection 3.2.3](#) for more insights on how to use cycleGAN for MR-harmonization.

Unfortunately, although this model shows great potential in harmonizing scans between two sites, it cannot be generalized to any unseen site or scanner. It must be trained for each pair of sites in the database, which makes the solution unsuitable for use in clinical practice.

- **DISENTANGLED LATENT SPACE REPRESENTATION** was first introduced by [Dewey et al. \(2020\)](#). It consists in representing both the anatomical and the contrast information in two distinct latent spaces. Then, a decoder takes it as input to reconstruct an image

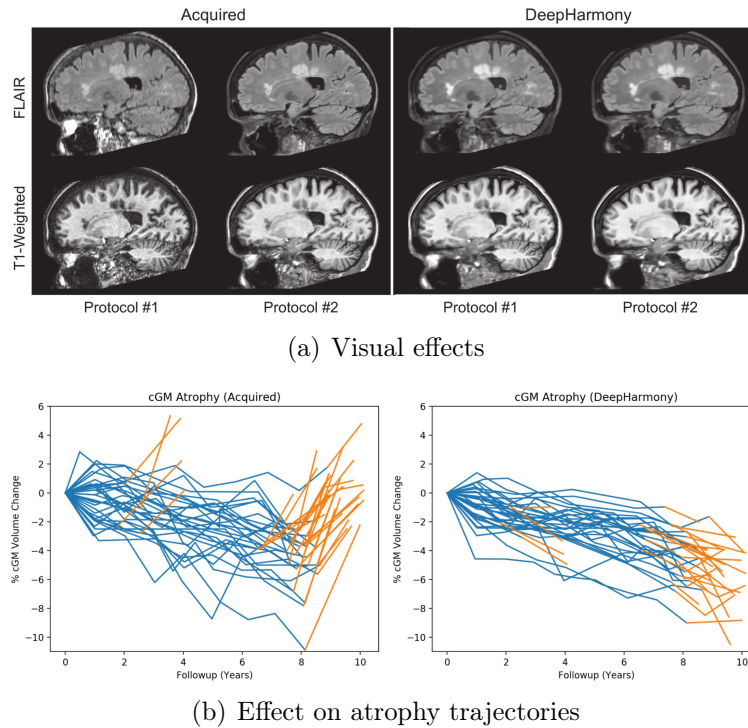


Figure 2.15 – DeepHarmony. **Top:** Representative sagittal slices for the same subject showing acquired images on the left and DeepHarmony-harmonized on the right. **Bottom:** Longitudinal trajectories for cortical gray matter (in % from baseline). Protocol1 is shown in blue and Protocol2 is shown in orange. Illustrations adapted from (Dewey et al., 2019).

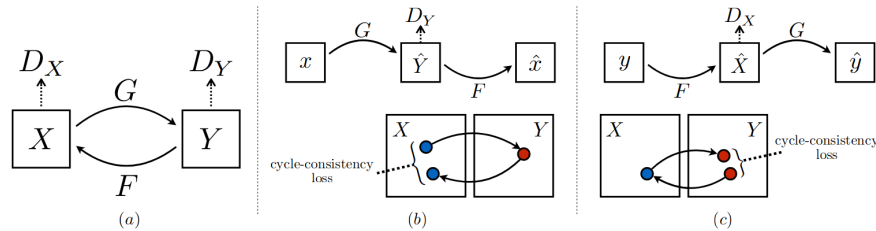


Figure 2.16 – CycleGAN. Illustrations representing CycleGAN architecture composed of two GAN networks consisting of two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X . D_Y encourages G to translate X into outputs indistinguishable from domain Y , and vice versa for D_X and F . Authors introduce two cycle consistency losses to ensure that if we translate from one domain to the other and back again we arrive where we started. Illustration adapted from (Zhu et al., 2018)

combining both encoded anatomical and contrast information. The authors trained 3 U-Net (Ronneberger et al., 2015) models for encoders (for contrast and anatomy) and decoder. The model requires 2 MR sequences from each subject (e.g. T1w and T2w sequences) from training. By using two contrasts with similar anatomies, the model can learn how to transfer contrast information while preserving the anatomy. Once the whole model is trained, one can apply the contrast of a chosen site to any anatomical information to perform image harmonization.

This solution was the first to propose such a disentangling procedure. It is based on

unsupervised learning, using the dual information of T1w and T2w scans to generate outputs from different combinations. This way, one can adapt the contrast of any scan while preserving the anatomy. It is indeed crucial for harmonization as the induced bias impacts mainly the contrast. However, even if the authors showed a positive effect of this method on contrast and similarity scores on traveling subjects data, this disentangled approach requires to be validated on larger datasets with a clinical application. Moreover, such solutions require the acquisition of both T1w and T2w sequences for each training subject, and will likely need fine tuning for each new site to be harmonized.

- **CALAMITY** is a deep generative network developed by [Zuo et al. \(2021b\)](#) that is able to modify MR-scan contrast while preserving the anatomy. It is based on the previous research study by [Dewey et al. \(2020\)](#) on disentangled latent space representation. In this study, the authors proposed to validate the disentangled approach using a new architecture. Their proposed network is made of:
 - a contrast encoder made from a variational encoder ([Kingma and Welling, 2014](#)).
 - a anatomical encoder made from a U-Net [Ronneberger et al. \(2015\)](#)
 - a discriminator connected to the anatomical encoder, ensuring that the encoder does not capture site-specific features
 - a U-Net decoder connected to both encoders output

The encoders learn a global representation of the data in their respective latent spaces. Once trained, the model is able to adapt any anatomical information to a given contrast. Its unified architecture allows it to harmonize as many domains (site or scanner) as there are in the training dataset.

The architecture proposed was also developed in a 2.5D fashion as proposed in [Dewey et al. \(2019\)](#). Instead of using the median, they use a 3D fusion convolutional neural network taking as input the reconstructed 3d outputs of each model trained along the three natural axis.

This model made a great improvement in the field of deep generative models for MR-harmonization by validating the disentangled approach proposed by [Dewey et al. \(2020\)](#). The results were very impressive as they could improve similarity score for different acquisitions among traveling subjects, and preserve the anatomical information of the original scan while removing domains related noises. However, the main downside of this solution is that it requires to be fine tuned every time a new domain is encountered. This downside by itself makes the solution hard to introduce in clinical practice as new domains are very likely to be met. Note also that its training requires T1w and T2w acquisitions for every subject, which can be limiting in some situations.

- **STARGAN** is an original DL model proposed by [Choi et al. \(2018\)](#). It allows simultaneous training of multiple datasets with different domains within a single network. A single generator is trained to transfer a given image into a given target domain. This architecture presents great advantages over the previous ones. Its unified model does not require paired images and is able to perform multi-site ($n_site > 2$) harmonization. It is also meant to generalize its training to unseen sites without the need for fine tuning. This solution has been used for MR-harmonization by [Bashyam et al. \(2021\)](#). Authors trained this model on 8,876 subjects from 6 sites. They validated StarGAN harmonization using an age prediction model. They could show a significant improvement

of age prediction metrics after harmonization. However, in their study, the model was trained and used on all present scans without using cross-validation which is a critical point for the following validation steps. Moreover, the study did not report any direct qualitative improvements (MAE, SSIM, PSNR, etc...) on traveling subjects data. In addition, such architecture seems less adapted to MR-harmonization than the previous two, as not disentangling anatomical and contrast information.

2.2.3 Unlearning Deep-Learning harmonization

Inspired by domain adaptation techniques, some research teams have proposed to address MR-harmonization directly at the level of clinical application. Instead of generating harmonized images, these solutions propose to estimate unwanted noises while addressing a clinical application (segmentation, classification, etc...). Usually this is done by integrating an additional unlearning module to the main model. This module is usually connected to the latent space representation of the initial data. It will act as a 'filter', forcing the encoder to produce scanner invariant representation. Another unlearning technique is called 'Pruning' and consists of removing neural connections inside the model to prevent it from over-fitting or learning domain related features.

At first sight, these solutions seem very interesting as they skip the classical harmonization step: 'Generating a corrected image'. This part usually requires a lot of data and models can be very hard to train (see [subsection 2.2.2](#)). On the other hand, as these models do not generate harmonized images, unlearning modules will have to be introduced and trained for every clinical application. On the contrary, generative harmonization models seek to produce images that can be used and reused once harmonized.

- **UNLEARNING OF DATASET BIAS** was first proposed by [Dinsdale et al. \(2021\)](#). Directly inspired by a domain adaptation technique called Domain Adversarial Neural Networks" proposed by [Ganin et al. \(2016\)](#) and illustrated in [Figure 2.17\(a\)](#). The main contribution consists in a generic framework that optimizes the feature representation for a label predictor related to the main clinical objective, and a domain classifier that aims to predict the source of the data. Eventually, the network tries to minimize the loss of the label predictor and to maximize the loss of the domain classifier. In an adversarial fashion, the feature extractor (usually a CNN encoder) will try to fool the domain classifier. As the number of domains is usually greater than one, it forces an uniform output by minimizing a confusion loss defined as bellow:

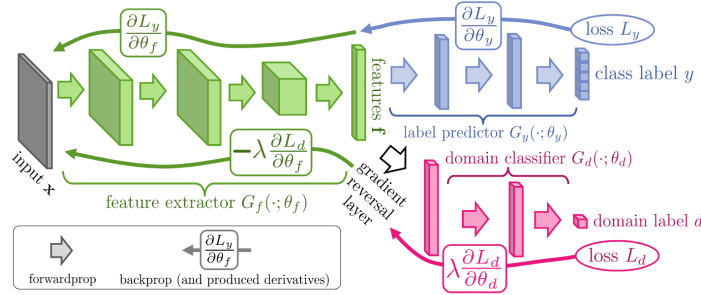
$$l_{confusion}(P) = -(1/N) \sum_{i=1}^N \sum_{s=1}^S \log(p_i^s)/S \quad (2.10)$$

With $P = [p_1; \dots; p_S]$ being the softmax output from the domain classifier, corresponding to the probability to belong to different domains (1, ..., S), Y is the ground truth domain affiliation vector, and N is the sample size.

The authors validated this framework using three different databases for a total of 8418 T1w scans. All results were obtained using a 5-fold cross validation procedure. They used two clinical experiments to evaluate the impact of such an unlearning module. First, as already done in previous harmonization studies [Bashyam et al. \(2021\)](#); [Pomponio et al. \(2020\)](#), they considered the task of brain age prediction from T1w scan. As shown in [Figure 2.17\(b\)](#), classical age prediction models could not generalize their

training to unseen databases. After introducing the unlearning module, MAE scores were much closer and the model could generalize its training to other datasets. Similarly, the authors showed a positive impact of such techniques for brain segmentation tasks (WM, GM and CSF in the study).

This study demonstrates the real benefit of introducing an unlearning module to any model when dealing with multi-centric datasets.



(a) Domain Adversarial Neural Network architecture

Training Data							
	B	O	W	Biobank MAE	OASIS MAE	Whitehall MAE	Scanner Classification Accuracy %
Normal Training							
1.	✓	×	×	3.25 ± 2.36	16.50 ± 6.77	13.81 ± 5.42	-
2.	×	✓	×	5.61 ± 3.52	4.27 ± 3.79	6.73 ± 4.82	-
3.	×	×	✓	5.61 ± 3.65	5.22 ± 4.83	3.15 ± 2.81	-
4.	✓	✓	×	3.30 ± 2.50	4.00 ± 2.78	4.71 ± 3.42	98 (50)
5.	✓	×	✓	3.31 ± 2.49	4.45 ± 3.53	3.05 ± 2.84	100 (50)
6.	×	✓	✓	5.71 ± 3.59	4.05 ± 3.71	3.21 ± 2.94	100 (50)
7.	✓	✓	✓	3.24 ± 2.47	4.19 ± 3.50	2.89 ± 2.70	96 (33)
Normal Training - Eq (1)							
8.	✓	✓	×	3.45 ± 2.63	3.99 ± 2.85	4.56 ± 3.38	100 (50)
9.	✓	×	✓	3.42 ± 2.60	4.19 ± 4.01	2.99 ± 2.69	100 (50)
10.	×	✓	✓	4.56 ± 3.05	4.08 ± 3.73	3.18 ± 2.93	100 (50)
11.	✓	✓	✓	3.55 ± 2.68	3.90 ± 3.53	2.62 ± 2.65	98 (33)
Unlearning							
12.	✓	✓	×	3.41 ± 2.04	3.79 ± 2.99	4.60 ± 3.47	48 (50)
13.	✓	×	✓	3.41 ± 2.58	4.07 ± 4.12	2.81 ± 2.57	52 (50)
14.	×	✓	✓	3.38 ± 2.64	3.91 ± 3.53	2.82 ± 2.65	50 (50)
15.	✓	✓	✓	3.38 ± 2.64	3.90 ± 3.53	2.56 ± 2.47	34 (33)

(b) Impact on age prediction

Figure 2.17 – Unlearning modules. **A)**: The proposed architecture by Ganin et al. (2016) including the feature extractor (green), the label predictor (blue) and the domain classifier (pink). **B)**: Results comparing unlearning to training the network in different combinations on the datasets. Mean absolute error is reported in years. B = Biobank, O = OASIS, W = Whitehall. In Exp 8-11, authors used class weights for domain classifier training so that it is not driven by the largest dataset. Illustrations taken from Ganin et al. (2016) and Dinsdale et al. (2021).

- **ATTENTION-GUIDED DEEP DOMAIN ADAPTATION** was proposed by Guan et al. (2021). Authors developed an attention-guided deep domain adaptation framework for multi-site MRI harmonization and applied it to automated brain disorder identification. Their solution is made of four main components, 1) a feature encoding model for MRI feature extraction, 2) an attention discovery module to locate disease-related regions in brain MRIs, 3) a disease classifier and 4) a domain discriminator encouraging the encoder to learn domain-invariant MRI features. Unlike the previously proposed unlearning framework, here the domain classifier loss is directly integrated in the global model loss, meaning that one domain is selected to be the ‘target’ during all training. However, instead of just aligning feature distributions of source and target domains as done

in [Dinsdale et al. \(2021\)](#), their method also aligns the attention maps learned from convolution layers.

Authors evaluated their proposed solution through 4 classification experiments related to AD: 1) AD classification; 2) MCI conversion prediction, which is a crucial diagnosis in Alzheimer disease routine; 3) AD vs. MCI classification and finally 4) MCI vs. control classification.

Authors reported impressive results on all experiments, and also compared their solution to other harmonization solutions. In all cases, they could outperform other methods, reaching the highest classification metrics. However, it is important to note that the authors did not report the actual effect of the domain adaptation component, as they did not present the results without this component. In addition, their model was designed to harmonize image into one selected ‘target’ domain, and this choice is likely to have an impact on the models performance.

- **MODEL PRUNING** is an original solution designed to prevent models from over-fitting. A primary cause of this over-fitting is the vast number of parameters in classical CNNs such as U-Net ([Ronneberger et al., 2015](#)). Introduced by [Molchanov et al. \(2017\)](#), pruning consists here in successively removing neural layers inside the model. Results showed that the parameters of the smallest magnitude generally have the least impact on the network’s output and this affects the choice of the layers that will be removed. [Dinsdale et al. \(2022\)](#) proposed to use such a technique in an MR-segmentation task. They iteratively trained for 1 epoch and pruned a U-Net model until the model performance was penalized or the desired model size was reached. Doing so, [Dinsdale et al. \(2022\)](#) showed that U-Net could better generalize its training to unseen datasets for similar segmentation tasks. This approach is really interesting as it reduces the over-fitting effects to training datasets but also reduces the number of parameters constituting the model, resulting in a lighter model.

2.2.4 Methods summary

We have seen that many harmonization solutions have been proposed over the years. We propose in [Table 2.1](#) to summarize these methods with their main features, pros and cons.

	Solution type	Preserves Local variations	Tested on clinical application	No need for traveling subjects	Fine-tuning for new clinical question	Fine-tuning for unseen sites	Max. number of target sites	Disentangling approach	Generative solution
Nyul et al. (2000)	Statistical	✗	✓	✓	—	✗	Undefined	✗	✓
Shinohara et al. (2014)	Statistical	✗	✓	✓	—	✗	Undefined	✗	✓
Fortin et al. (2016)	Statistical	✗	✓	✓	—	✓	N = 2	✗	✓
Fortin et al. (2017)	Statistical	✓	✓	✓	—	✓	n training	✗	depends
Pomponio et al. (2020)	Statistical	✓	✓	✓	—	✓	n training	✗	✗
Da-ano et al. (2020)	Statistical	✓	✓	✓	—	✓	n training	✗	✗
Beer et al. (2020)	Statistical	✓	✓	✓	—	✓	n training	✗	✗
Zhu et al. (2018)	Deep-Learning	✓	✓	✓	—	✓	N = 2	✗	✓
Dewey et al. (2019)	Deep-Learning	✓	—	✓	—	✓	N = 2	✗	✓
Dewey et al. (2020); Zuo et al. (2021a)	Deep-Learning	✓	✓	✓	—	✓	n training	✓	✓
Bashyam et al. (2021)	Deep-Learning	✓	✓	✓	—	✗	Undefined	✗	✓
Dinsdale et al. (2020), Guan et al. (2021)	Deep-Learning	✓	—	✓	✗	—	—	✗	✗

Table 2.1 – Harmonization methods summary. ✓ means that the solution verifies the condition while ✗ means the opposite. In addition, the green color signifies that a verified condition is a positive argument for the harmonization solution.

Chapter 3

ComBat versus cycleGAN for two sites MR images harmonization

Abstract

In this chapter, we present the effect of two harmonization methods, ComBAT ([section 2.2.1](#)) and cycleGAN ([subsection 2.2.2](#)) when used to harmonize images from two acquisition sites. Both solutions were chosen as they seem, according to the literature, adapted for MRI harmonization but were never compared in that context. Once the images harmonized, it is also difficult to evaluate the benefits for further clinical applications. Through five experiments, performed on synthetic and real data, we propose here to benchmark these two methods. Focusing on T1-weighted MR images, we investigate the effects of each harmonization approach using radiomic features (meant to represent different aspects and properties of images) and Support Vector Machine (SVM). The study reports that both methods perform well for removing various types of noises while preserving manually added synthetic lesions. They also seem to be adapted for removing site effects on data coming from 2 different sites while preserving biological information. Moreover, while each method improves autism data classification, they have different impacts on radiomic features and appear complementary in several aspects. This work was presented as a digital poster during the [2021-AI4Health winter school](#). It was also accepted for an oral presentation at the national French conference SFRMBM-2020. Unfortunately due to Covid restrictions the event was canceled.

Keywords— Brain, MRI, harmonization, deep-learning, radiomic features, ASD classification

CONTENTS

3.1	Introduction	35
3.2	Materials and methods	35
3.2.1	Data	35
3.2.2	The Combined Association Test: ComBat	35
3.2.3	cycleGAN	36
3.2.4	Experiments	38
3.3	Results	41
3.4	Discussion	45
3.5	Conclusion	46

3.1 Introduction

A general introduction to MRI data harmonization was presented in [section 1.2](#). In this study, we propose to compare 2 promising methods for the harmonization of T1-weighted MRIs. ComBat, a statistical method proposed by [Fortin et al. \(2017\)](#) for DTI harmonization, and cycleGAN, a deep-learning model introduced by [Zhu et al. \(2018\)](#). In order to describe the effects of both approaches comprehensively, we ran five different experiments performed on synthetic data as well as real *in vivo* images. We first assessed the capacity of the 2 methods to remove manually added global noises in the images as well as the ability to preserve manually added lesions. We also investigated harmonization’s benefits on image analysis like site classification and Autistic Syndrome Disorder (ASD) detection. Harmonization effects were evaluated using radiomic features, known to be sensitive to harmonization ([Da-ano et al., 2020](#); [Orlhac et al., 2019](#)). Finally, we evaluated the impact of harmonization solutions on different radiomic features groups. These derived metrics are meant to represent main features of images, like shape, contrast or texture. It is likely that induced site biases will affect each family of features differently, and so will the harmonization methods.

3.2 Materials and methods

3.2.1 Data

As for all experiments in this PhD thesis, the open access [ABIDE database](#) was used. It is a multi-center project led in 2014 by [Di Martino et al. \(2014\)](#), focusing on autism disorders among children. It gathers more than 800 pediatric autistic patients and controls. In this study, we used healthy 3DT1-MRI scans from 2 different sites A (GU) and B (OHSU), selected as the ones presenting the largest collection of subjects. Age range was from 8 to 14 years old for both sites with similar sex distribution. All acquisitions were realized on two different 3T Siemens TIM trio scanners (each one coming from one site). MR images were first co-registered to age specific 152-MNI templates publicly available ([Sanchez et al., 2012a](#)). The brain was then extracted using Robex ([Iglesias et al., 2011](#)) and N4Bias ([Tustison et al., 2010](#)) was used to correct for inhomogeneities of intensity. After a manual quality check, we removed 11 scans presenting either acquisition artifacts or brain extraction issues. Finally, 51 scans were extracted for site A (56 for site B). Data was eventually re-scaled between $[-1;1]$.

3.2.2 The Combined Association Test: ComBat

As presented in [section 2.2.1](#), as a result of two studies ([Fortin et al., 2018, 2017](#)), ComBAT has quickly been considered as a gold standard harmonization solution. Although relying on a strong hypothesis for parameters prior distributions, ComBat is known to be robust to small sample sizes and is easy and fast to use. Later studies, like [Orlhac et al. \(2019\)](#), showed ComBAT’s efficiency for harmonizing radiomic features derived from PET, another imaging modality.

In this study we used the open-access python module ‘[NeuroCombat](#)’ (2020 version, deprecated today) developed by [Fortin et al. \(2018\)](#).

3.2.3 cycleGAN

As introduced in [subsection 2.2.2](#), cycleGAN is a deep-learning model that can resolve Image-to-Image translation tasks. It was initially developed by [Zhu et al. \(2018\)](#). The principle relies on two GANs learning how to map images translation in opposite order, the architecture is represented in [Figure 3.1](#).

A more recent study ([Modanwal et al., 2020](#)) reports the great potential of such a bidirectional model to overcome multi-centric induced bias for MR breast scan harmonization.

We used cycleGAN in a 2D fashion, working on axial slices. As these models are known to require a large amount of data for a successful training, and to avoid over-fitting, using 2D images is a way to increase our sample size and to reduce the number of parameters to fit (2D models are less complex than their 3D equivalent ones).

Each GAN has a Pix2Pix ([Isola et al., 2018](#)) architecture, consisting of an UNet ([Ronneberger et al., 2015](#)) generator and a patchGAN as a discriminator. The choice of the discriminator's field of view size was motivated by the results obtained by [Modanwal et al. \(2020\)](#). We used the LeakyReLU activation function for the encoder part of the generators and discriminators. Classical ReLU function was used for the decoder part of the generators. Downsampling (resp. upsampling) was done through convolutional (resp. transposed-convolutional) layers. Finally the model loss was composed of classical binary cross-entropy loss (l_{disc}) (used to train discriminators), a l1-cycle loss consistency (l_{cycle}) and a l1-loss (l_1) between the input and output of each generator. This final term was found to be helpful for training and led to better convergence.

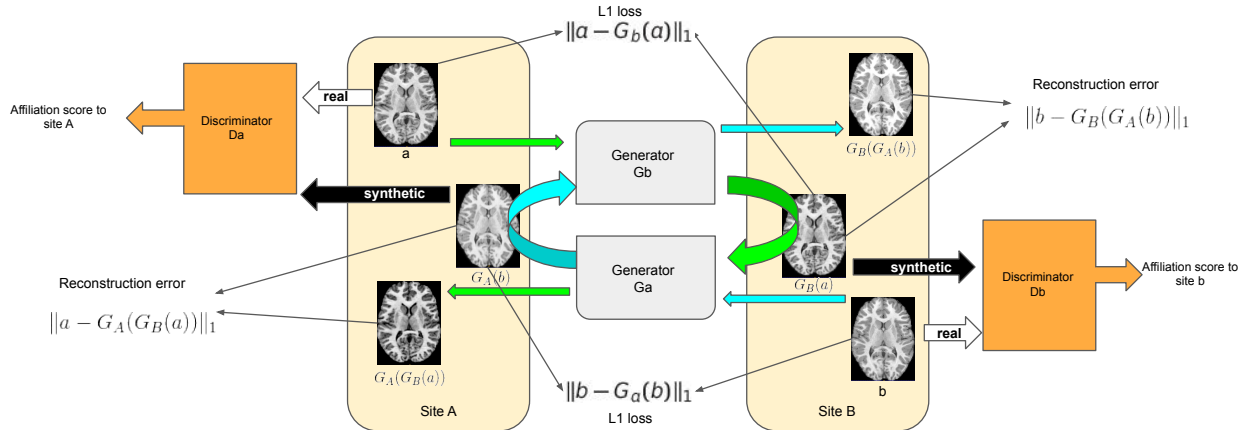


Figure 3.1 – CycleGAN architecture used for MRI data harmonization in this study

Note that the model presented above differs from the original cycleGAN architecture. Based on latest publications ([Isola et al., 2018](#); [Yi et al., 2018](#)), we use U-Net as a generator instead of a modified Resnet. This type of implementation is commonly used: for instance, on TensorFlow's website, the proposed implementation in the [cycleGAN tutorial](#) contains U-Nets as generators. This U-Net model has become a standard for medical image segmentation and generation. Furthermore, the skip-connections between the encoder and decoder layers allow the model to better preserve image anatomy, which is a crucial aspect in MRI data harmonization.

In addition, we added the l_1 term to prevent the model from well-known instabilities leading

to drastic changes to the image. This term is also helpful for training as these kinds of models are known to be difficult to train. The order of magnitude of this l_1 is however very small (10^{-4}) compared to the global training loss (its order of magnitude was 10). Thus, we do not expect a major impact on training results.

$$l_{cycleGAN} = \lambda_1 l_{disc} + \lambda_2 l_{cycle} + \lambda_3 l_1 \quad (3.1)$$

With, $\lambda_1 = -1$; $\lambda_2 = 100$; $\lambda_3 = 3$; $l_{cycle} = \|a - G_a(G_b(a))\|_1 + \|b - G_b(G_a(b))\|_1$; $l_{disc} = (\log(1 - D_a(G_a(b))) + \log(1 - D_b(G_b(a))))$; $l_1 = \|a - G_b(a)\|_1$.

The model was trained through 500 epochs, using a batch size of 8. Learning rate was initialized to $6 \cdot 10^{-4}$ and then reduced (model independently) on validation loss plateau by a factor 0.8. All training was done using Tensorflow 2.0 on a Quadro P2000 GPU / Intel Core i7-8700K CPU. Each training session took about 2 hours.

For all experiments, we used 10-fold cross-validation to train and infer all control subjects. To infer subjects with ASD, we trained a model on all control subjects and used it on patients' data.

Algorithm 1 presents the workflow used for cycleGAN for all experiments. Note that cycleGAN is used to harmonize data between 2 sites. At the time of the inference, we only use one generator (G_b) to harmonize only the data from site A into the domain of site B.

Algorithm 1 CycleGAN overall workflow: Training-Validation-Inference protocol

Gather same sequence MRIs from 2 sites : OHSU & GU sites from the ABIDE database

#Preprocessing steps

Brain extraction with Robex algorithm

Co-registration to age-specific MNI templates

Intensity Bias field correction using N4 Bias algorithm

Visual quality check : brain extraction & image acquisition

Rescaling data between $[-1; 1]$

Extracting 2D axial slices while removing background slices

#Training steps → Inferring control subjects

Splitting control data in 10 folds for cross validation

for $i = 0; i < 10; i++$ **do**

Train cycleGAN using $(F_i : F_{(8+i)\%11})$ as training sets, $F_{(9+i)\%11}$ as validation set, and $F_{(10+i)\%11}$ as test/inference set.

end for

Second training phase → Inferring ASD subjects

Gather all control subjects in one fold and ASD ones in another

Select randomly 10 control subjects for validation steps

Train cycleGAN on training control subjects

Once the model trained, run inference on ASD subjects

3.2.4 Experiments

To compare the 2 harmonization methods, we ran 3 experiments on synthetic data and 2 on real *in vivo* data (detailed in next sections). Synthetic data were used to assess the ability to remove global noise or preserve known local structures or local biological variations. On the other hand, real data were used to estimate methods efficiency to remove site effects, and their ability to improve further clinical analyses.

CycleGAN was trained from scratch for each experiment. For all classification tasks, we used SVM with a radial basis function kernel to classify data before and after harmonization. To evaluate the specificity and sensitivity of our classifier, we used the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. As visual inspection is not sufficient to evaluate the effect of harmonization on the images, we extracted radiomic features, known to be sensitive to site effects (Orlhac et al., 2019). We used the pyradiomics python API developed by van Griethuysen et al. (2017) to extract 101 features. These features aim to represent different aspects of MRI images such as shape, contrast or texture. Radiomic features are organized in families which are described on the API's website. In all cases, we first selected the 'most correlated features' using Pearson tests (ran independently for each feature) with the characteristic of interest (site affiliation, presence of added noise, etc...) using 10^{-3} as p-value threshold. This step was essential to focus on the effects of the methods on characteristics of interest only. We also ran Pearson tests after harmonization on previously selected radiomic features to better understand the impact of both methods on these features. Finally, we investigated correlations between radiomic features and biological ones (sex and age). Our hypothesis was that harmonization should increase or at least preserve correlations when existing.

All classification procedures were done using a 8-folds cross-validation (while a 10-fold cross-validation was used to train our cycleGAN model) repeated 10 times in order to have various sets combinations and to be able to compute statistical evaluations on the performance of the classifiers. To visualize the results, we reduced the dimension with PCA and TSNE (Maaten and Hinton, 2008). PCA was first used to assure orthogonal representation of our data (8 components used, representing around 95% of total variance), and then TSNE to visualize our data along 2 axes. Once dimensions were reduced, it was possible to observe two clusters of points corresponding to different sites or data types. For validation, we only used PCA-reduced data, using the 8 first principal components (representing more than 95% of the total variance), as there was no need for data visualization using the TSNE algorithm. Finally Welch's t-tests (Welch, 1947) were run to validate if results were statistically significant or not. We ran these tests on every combination of data under the null hypothesis "method does not impact SVM accuracy" and "both methods have same performances". We then observed the p-values of these tests and rejected the H_0 if $p < 0.05$. Because variances of the results obtained by the two methods could not be considered as equal, we used the Welch t-test to compare whether the differences observed were statistically significant.

Experiments 1-2: ability to remove synthetic global noises

This first experiment evaluated the ability of both methods to remove synthetic global noises added manually to the images. First, a synthetic global Gaussian intensity shifts was added, centered in the middle of images in order to simulate variations in RF coil homogeneity. This added Gaussian noise follows a 2D $N(0, 0.3^2)$ distribution. It was then multiplied by a

factor 1.6 so that its intensity in the middle of the images was increased by 60%. Secondly, a classical Gaussian noise was added to induce multiple artifacts and to reduce contrast in the images. This was done by adding independently $\epsilon \sim N(0, 0.6^2)$ to every voxel. [Figure 3.2](#) illustrates the global noises added during these experiments.

Here, both methods were fitted using all the data from site A, including the ones altered with synthetic global noises. We considered the presence of synthetic noise as induced site noise and then used cycleGAN and ComBAT to harmonize modified and original data together. Doing so, we could consider two different domains of images, the original ones and the altered ones. The objective was then to use both methods to learn how to remove these noises. To better stick to a realistic situation, we did not train the cycleGAN in a supervised way for noise reduction. In fact we could have added an additional loss (between the output and the original data) to force it to remove the known added signal. However, as the idea was to simulate an unknown global signal induced by scanners, we chose to stay in real life conditions.

As presented in [Figure 3.3](#), the SVM was used here to classify the presence of added global noise in the considered images. As the harmonization solutions aim to remove the global added signal, we expect to see a reduction of the SVM accuracy metrics on the processed data compared to the original ones.

Experiment 3: ability to preserve synthetic lesions

Contrary to global variations, local variations are most likely of biological origin. Therefore, harmonization must preserve these local variations as we do not want to lose any biological information in the process.

To assess if local changes in image intensities were retained after harmonization, a synthetic localized spherical Gaussian intensity shift was added to some randomly sampled data. This mimics hyperintensities that can be found in patients presenting several pathologies (e.g. AD, gliomas or tumors) or among subjects with stroke or trauma. For this experiment, cycleGAN and ComBAT were trained on unmodified healthy data, and then used on the whole dataset (original + altered data). As for global noise removal, we followed the same workflow presented in [Figure 3.3](#). Hyper-intensities preservation was estimated by SVM classification accuracy before and after harmonization. To verify that harmonization improves, or at least preserves, synthetic lesions classification, we used SVM to classify the presence of synthetic lesions. This experiment was run several times with different ‘lesion’ radius, so that we could evaluate the impact of harmonization with respect to the size of local variations. Radiomic features extracted on the all brain were used as input for our classifier. Moreover, we computed first order statistics (mean and variance) in the altered regions to probe possible geometrical modifications due to the harmonization process.

Experiment 4: Site effect removal

One of the key points of data harmonization is the removal of site or scanner induced bias. These biases often lead to easy discrimination of the origin of the data and it is usually possible to predict the origin of each scan in multi-centric datasets ([Fortin et al., 2017](#); [Liu et al., 2020](#)). This demonstrates a domain specific signal in the data and the need for data harmonization.

As in previous experiments, we evaluated data harmonization of the 2 selected sites through

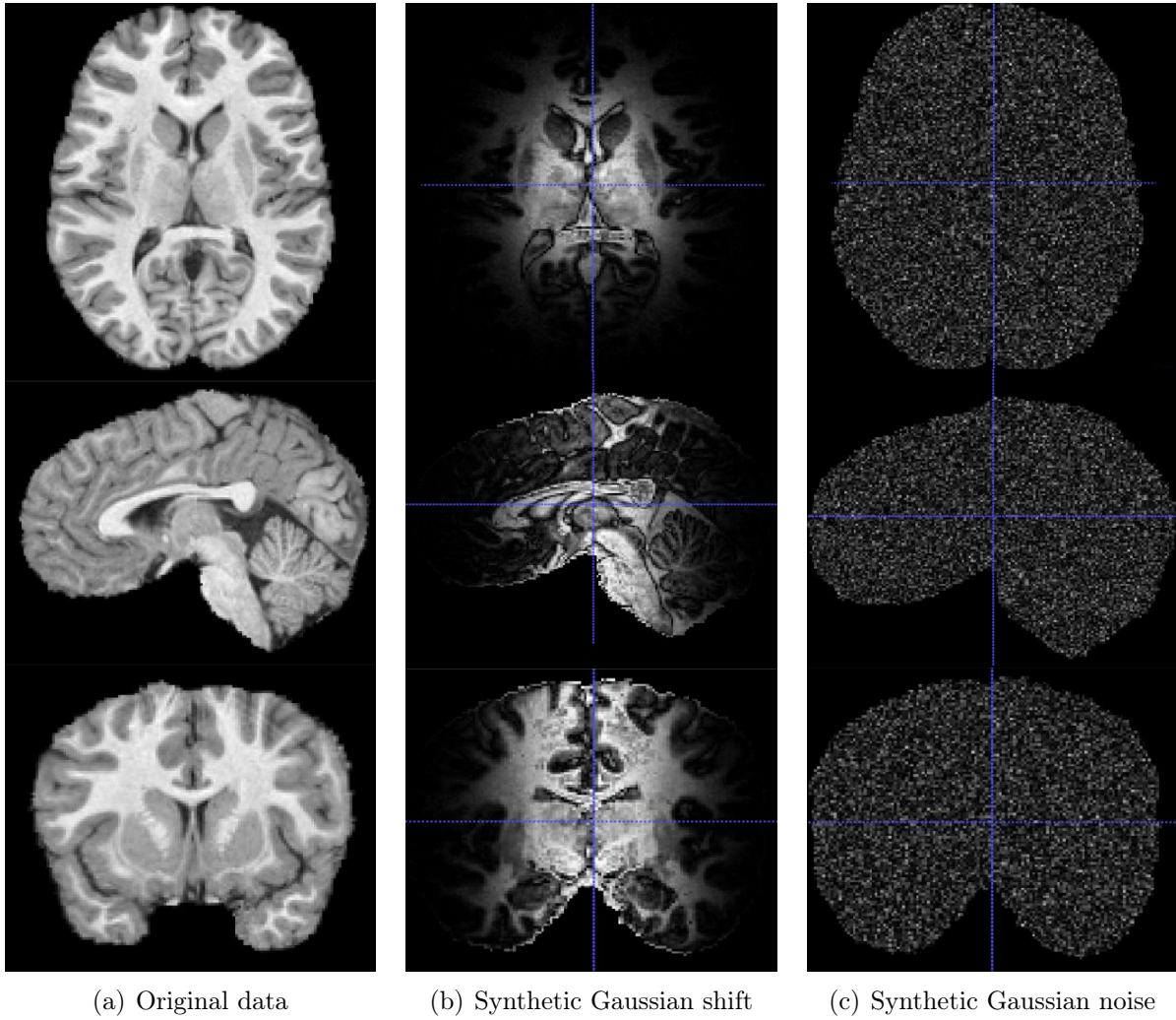


Figure 3.2 – Illustration of the synthetic global noises added for experiments 1&2. **A)** Original scan; **B)** Global 2d Gaussian shift centered in the middle of each slice; **C)** Global Gaussian noise.

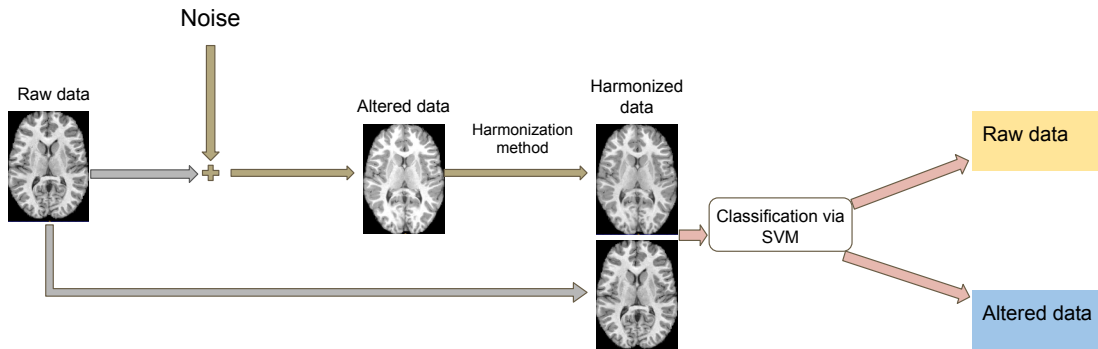


Figure 3.3 – Synthetic Noise harmonization workflow.

SVM accuracy metrics. The hypothesis here was that no classifier should be able to accurately detect data origin. We thus tried to classify data origin before and after harmonization

to see how harmonization could penalize the classifier.

To estimate the impact of harmonization on image features, we also ran Pearson tests with all radiomic features independently for site affiliation and age correlation. For this task, the hypothesis was that harmonization should preserve or enhance correlation between radiomic and biological (age here) features. On the other hand, harmonization should reduce correlation between radiomic features and site correlation. To evaluate data harmonization effect, we compared the numbers of significantly ($p < 5E10^{-2}$) correlated features with site affiliation before and after harmonization (expected to decrease). We also compared the same numbers for features correlated with age (expected to increase).

Experiment 5: ASD patients classification

This last experiment focuses on the second main aspect of data harmonization: ‘it must preserve biological information’ in order to improve further analyses. This part is often hard to quantify, and there is no reference protocol proposed in the literature.

In this study, we ran a clinical classification task (ASD patients vs healthy controls) on data from site A and B to evaluate if SVM performs better on harmonized data than on raw data. Similarly to [section 3.2.4](#), cycleGAN was trained on control data from both sites. Inference was done on all subjects from site B using a 10-fold cross-validation procedure. For this experiment, radiomic features were extracted on the whole brain.

3.3 Results

[Table 3.1](#) (exp1-2) shows that ComBAT performs well on removing simple global noises (Gaussian intensity shift and Gaussian noise) while cycleGAN does not remove these noises correctly. SVM AUC metrics drop from 1 to 0.44 ± 0.02 after ComBAT while remaining close to 1 after cycleGAN. [Figure 3.4](#) illustrates the effect of each method on the global noises considered.

On the other hand, [Table 3.2](#) presents SVM AUC metrics on synthetic lesions (local noises) classification when the lesions size (here its radius) varies. It shows that, in all cases, each method does not penalize the AUC of the SVM. Furthermore, for small synthetic lesions, we cannot assure a benefit from both methods as they could not be detected before nor after harmonization. For larger lesions radius, ($\geq 24mm$) it is clear that SVM performs better on harmonized data. [Table 3.2](#) also shows that in these cases (larger radius), cycleGAN better improves SVM performance, reaching an AUC of 1 for a radius larger than $32mm$. Note that we want an AUC as close to 1 as possible as we want the classifier to detect local variations after harmonization.

[Figure 3.5](#) illustrates the impacts of both methods on one original image. Both outputs seem very realistic, which is necessary, but insufficient. One can also observe a difference in both corrections, where ComBAT impact seems less homogeneous than cycleGAN and seems to have detected unwanted noise on the bottom of the brain. Similarly [Figure 3.6](#) presents harmonization impact on intensity distributions between the 2 sites considered. The hypothesis here was that harmonization should align both intensity histograms. Note that we do not seek a perfect alignment here as the two site populations are different. We can see on [Figure 3.6](#) that it was the case. While there was a clear shift between both histograms before

Experiment	Noise	Raw data	After ComBAT	After cycleGAN
1	Global Intensity Gaussian shift	1	0.43	1
2	Global Gaussian noise	0.96	0.46	0.85

Table 3.1 – SVM AUC on testing data for synthetic global noises classification (Exp1-2). In bold, the best performances (here the smallest values, as we seek a total noise removal after harmonization).

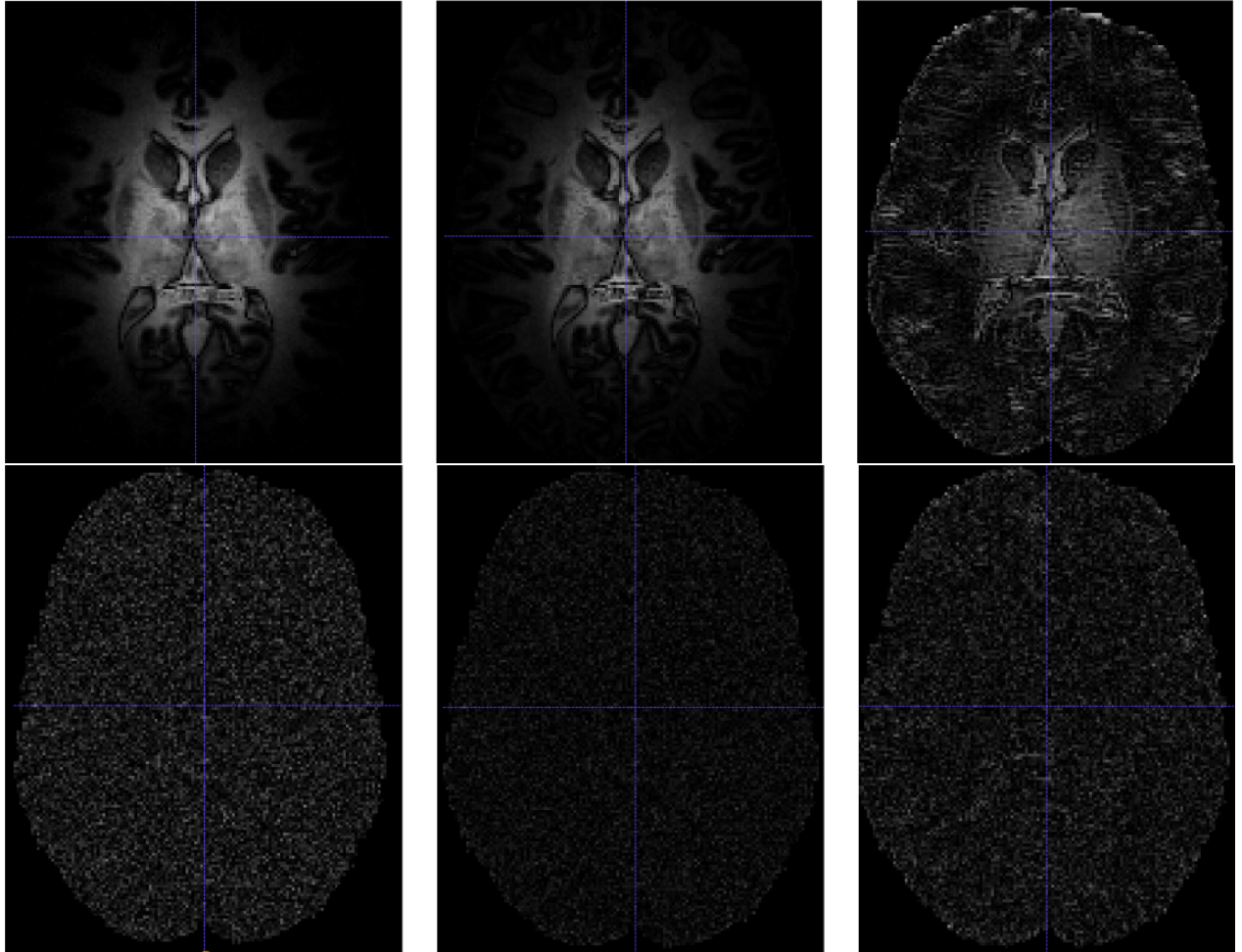


Figure 3.4 – Methods' effects on global noises during exp 1-2 ([section 3.2.4](#)). Results for global Gaussian intensity shift (exp. 1) are represented in the top row, and global Gaussian noise (exp. 2) in bottom row. From left to right: noise added to the image; absolute difference between original scan and comBat-denoised image; absolute difference between original scan and cycleGAN-denoised image.

harmonization, it was not the case anymore after ComBAT and cycleGAN.

For experiment 4 ([section 3.2.4](#)), [Figure 3.7](#) demonstrates the need for site harmonization as we can easily observe two clusters representing both sites when classifying the raw data

Synthetic lesion radius	Raw data	After ComBAT	After cycleGAN
8	0.50	0.50	0.50
16	0.50	0.50	0.50
24	0.50	0.57	0.70
32	0.83	0.75	1
40	1	1	1

Table 3.2 – Test SVM AUC evolution for synthetic lesions classification (Exp3). In bold, the best performances (here the largest ones as we want to detect local variations as well as possible).

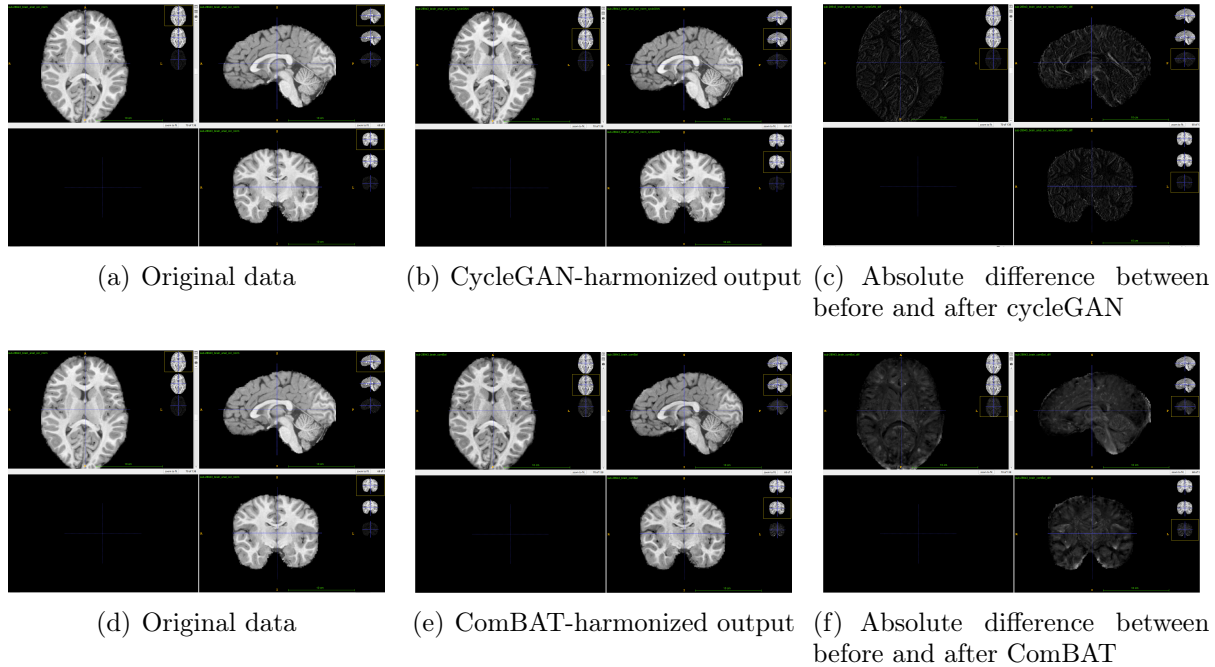


Figure 3.5 – Impact of both methods on one scan. **A&D)** Original data before data harmonization; **B)** ComBAT harmonized result; **C)** differential image showing induced modifications by ComBAT; **E)** cycleGAN harmonized result; **C)** differential image showing induced modifications by cycleGAN

(left). These clusters vanish after data harmonization from both methods (all data-points are then confounded in one same cluster). Moreover, using SVM AUC metrics, we see in both cases that the AUC drops, reaching a minimal value of 0.58 (resp. 0.68) after cycleGAN (resp. ComBAT) harmonization.

Finally, for experiment 5 (section 3.2.4), Figure 3.8 confirms the positive impact of both methods. The SVM achieves better performance on patient classification on harmonized data. ComBAT can increase the AUC metric from 0.67 to 0.74, while it reaches an AUC score of 0.89 after cycleGAN harmonization.

Additionally, Table 3.3 confirms these results as for both experiments on site and ASD classification, we have a significant improvement of SVM AUC after both methods. Moreover,

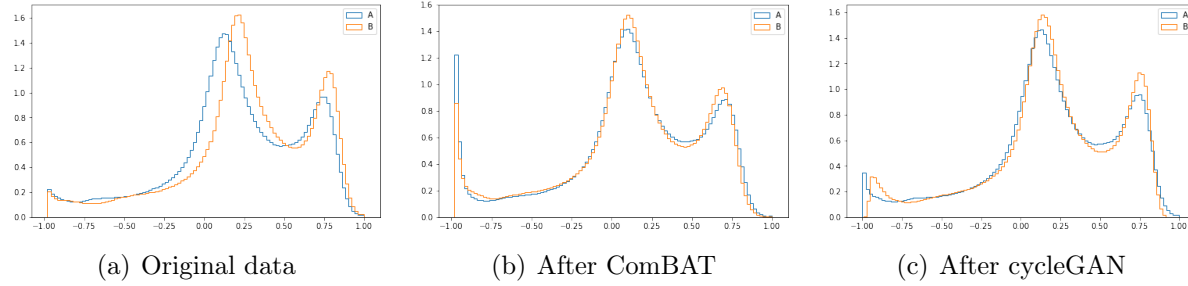


Figure 3.6 – Intensity distribution across sites before and after both methods. **A)** Original data before harmonization; **B)** after ComBAT; **C)** after cycleGAN

while ComBAT and cycleGAN show similar performances for site classification, we can observe better results after cycleGAN for ASD classification. The bottom right cell of Table 3.3 indicates that AUC scores obtained after cycleGAN harmonization are significantly higher than the ones obtained after ComBAT.

Additionally, Table 3.4 shows that each method significantly reduces the number of features correlated to site affiliation, while increasing the number of correlated features with age. This result is in agreement with the previous ones, highlighting the positive impact of both methods. Interestingly, even if both solutions have a positive impact on age and sex correlation, one can observe that ComBAT reduces the correlation to site more than cycleGAN, while this latter can better enhance features correlation with age. Another interesting point to mention is that the two methods do not impact the same features, as presented in the last column of Table 3.4. This result suggests that each method affects specific images' aspects and that they could complete one another. This final point is in line with all previously mentioned results as both methods performed differently for all experiments.

Classification	original data vs. ComBAT	original data vs. cycleGAN	ComBAT vs. cycleGAN
Site	$6.5 * 10^{-5}$	$2.4 * 10^{-5}$	0.13
ASD	$9.4 * 10^{-2}$	$7 * 10^{-3}$	$4 * 10^{-2}$

Table 3.3 – P-values of Welch's t-test comparing SVM performances on different types of data (Exp4 and Exp5). In bold significant differences of performance.

	Original data	After ComBAT	After cycleGAN	common features
Site	71	6	19	3
Age	27	49	34	22

Table 3.4 – The number of radiomic features significantly correlated with site affiliation and age, among the 101 extracted and obtained with a Pearson's test. The last column corresponds to the number of significantly correlated features common to both methods.

3.4 Discussion

Our results strongly support the need for data harmonization and show the efficiency of ComBAT and cycleGAN to tackle multi-center MRI study issues. This study shows that both methods reduce global added noises and site effects, while retaining local modifications, and improving the accuracy of the SVM for classifying synthetic lesions and ASD patients. However, it is important to point out differences between both methods' performances. While ComBAT seems to be more adapted to remove global noises and to improve correlation between radiomic features and site affiliation or age, cycleGAN shows better results at preserving local modifications and at improving statistical analysis of clinical studies. Note that the added l_1 term in cycleGAN training loss 3.1 could contribute to its lower performances on global noises as it forces the model not to modify the image too much. On the opposite this term is likely to explain the preservation of local hyper-intensities.

The last mentioned results about the differences between numbers of significantly correlated features in Table 3.4 is interesting. It could be explained by the fact that the ComBAT algorithm is built to remove site affiliation effects while preserving correlations with age (as it takes age and site affiliation as inputs). On the other hand, cycleGAN only takes MR images as input. It might be interesting to add other biological inputs like age and sex to the network to see how this could affect the results of Table 3.4.

While giving a closer look to Pearson's tests on radiomic features, we found that both methods preserve shape-related features, as expected. The opposite would have been problematic indeed, as site related noises do not alter images anatomy but mainly impact their contrast. The impacts of the two methods on other features families were found to be complementary: ComBAT performed well on GLRLM¹ features while GLSZM². In this study, we presented a workflow to evaluate harmonization techniques. We showed the importance of data harmonization when dealing with data from at least 2 centers. Indeed, we were able to precisely distinguish data from two acquisition sites, even though they both used the same type of scanners. We used our workflow to investigate the performances of two harmonization algorithms for anatomical MRI multi-center studies, ComBAT and cycleGAN. The two approaches could effectively remove unwanted site effects while preserving biological information. Both showed positive impact in all investigated experiments, as expected they could improve classification metrics for ASD patients classification. In this specific case, cycleGAN led to better results than ComBAT while the latter could better reduce global noises in images.

CycleGAN results demonstrated that a deep-learning method (non linear) was well-suited for harmonization and could outperform state of the art statistical methods (linear) such as ComBAT in certain conditions. We could have expected that the former outperformed the latter. Surprisingly, we also showed that this was not always the case. The two methods appear complementary in several aspects and had not the same effects on radiomic features. This could determine the choice of the techniques depending on the goal to achieve. Additionally, this opens a pathway to new solutions able to take advantages of both presented methods. One could think about combining both solutions for example of just developing ComBAT versions adapted to the clinical data. This has since been investigated in several studies which proposed upgraded ComBAT versions like comBAT-GAM, B&M-comBAT and the longitudinal-comBAT presented in section 2.2.1.

1. Gray Level Run Length Matrix
2. Gray Level Size Zone Matrix

Moreover, the fact that both solutions can not be used on new sites emphasizes the need of a more suitable solution. Otherwise, it is very unlikely that any harmonization solution will be used in clinical practice.

This last point connects directly to [chapter 4](#) in which we present an original solution designed to take over various drawbacks already mentioned. ones benefited better from cycleGAN harmonization. Other families were similarly impacted by both algorithms. An interesting point would thus be to investigate combinations of both methods (e.g. feed cycleGAN with data harmonized by ComBAT, or the other way around).

Another point worth mentioning is that ComBAT harmonizes all data when cycleGAN only modifies data from one site. This has an impact on our experiments. If transformations induce a noise while harmonizing, we should favor the one less impacting the images, here cycleGAN. Moreover, the ComBAT algorithm relies on a strong prior hypothesis for modeling voxels intensities, when cycleGAN tries to map the parameters directly without priors. CycleGAN also requires a much bigger sample size to be trained than ComBAT. [Fortin et al. \(2017\)](#) pointed out that ComBAT performs well even on small sample sizes. To illustrate these last points, we also ran experiment 4 ([section 3.2.4](#)) on sites A and B using 20 control subjects only. We found that cycleGAN was limited by the sample size and was not able to correct for site effects while ComBAT presented similar results as with the full dataset. Finally, we can point out that for each method, a new model has to be fitted for every new site encountered. This can be very time consuming and redundant, especially for cycleGAN which takes longer to be fitted than ComBAT. Thus, it could be very useful to investigate a way to generalize cycleGAN and ComBAT harmonization to every site and look for predictable features or biomarkers directly impacted by site or scanner noises.

3.5 Conclusion

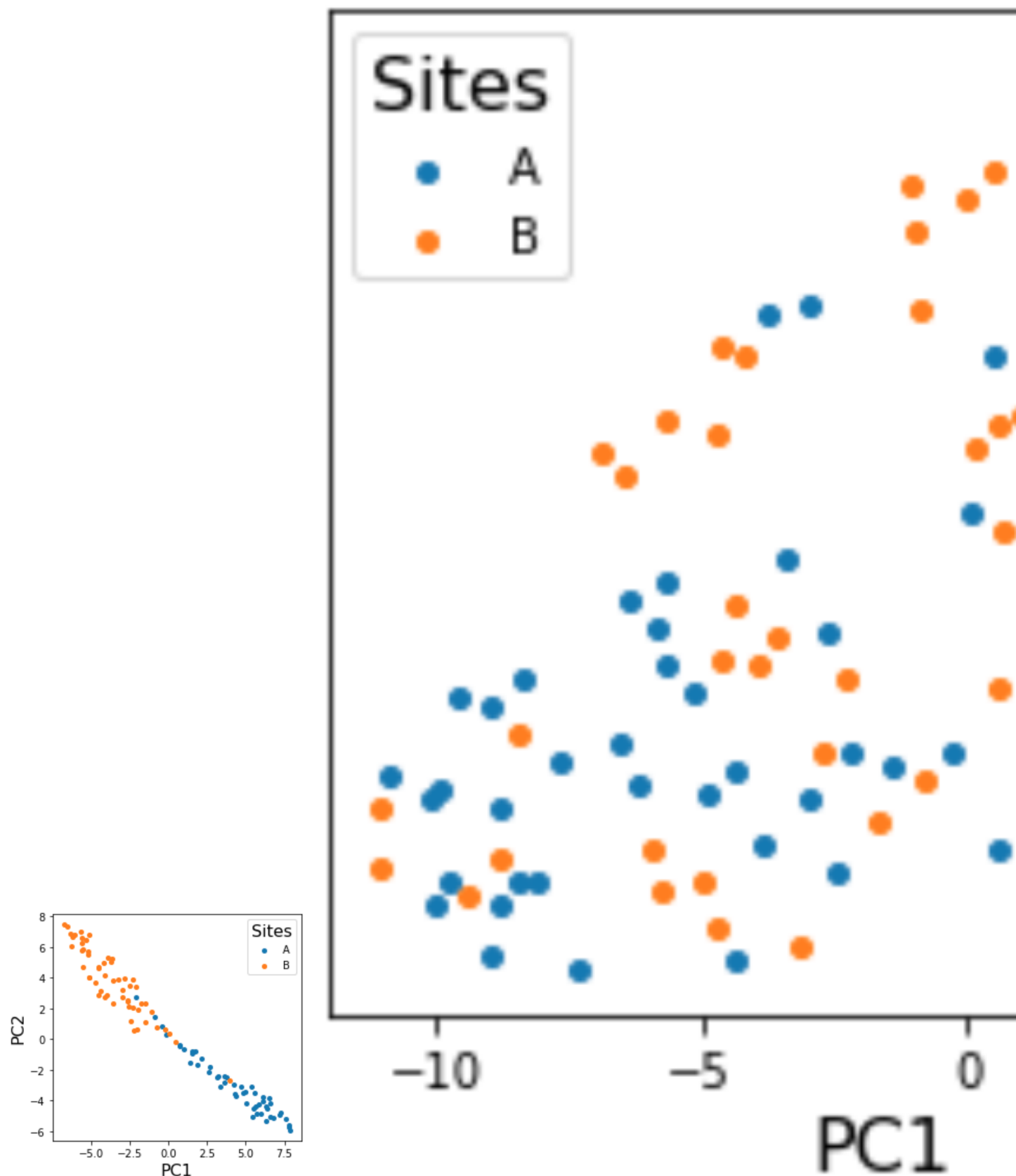
In this study, we presented a workflow to evaluate harmonization techniques. We showed the importance of data harmonization when dealing with data from at least 2 centers. Indeed, we were able to precisely distinguish data from two acquisition sites, even though they both used the same type of scanners. We used our workflow to investigate the performances of two harmonization algorithms for anatomical MRI multi-center studies, ComBAT and cycleGAN. The two approaches could effectively remove unwanted site effects while preserving biological information. Both showed positive impact in all investigated experiments, as expected they could improve classification metrics for ASD patients classification. In this specific case, cycleGAN led to better results than ComBAT while the latter could better reduce global noises in images.

CycleGAN results demonstrated that a deep-learning method (non linear) was well-suited for harmonization and could outperform state of the art statistical methods (linear) such as ComBAT in certain conditions. We could have expected that the former outperformed the latter. Surprisingly, we also showed that this was not always the case. The two methods appear complementary in several aspects and had not the same effects on radiomic features. This could determine the choice of the techniques depending on the goal to achieve. Additionally, this opens a pathway to new solutions able to take advantage of both presented methods. One could think about combining both solutions for example of just developing ComBAT versions adapted to the clinical data. This has since been investigated in several

studies which proposed upgraded ComBAT versions like ComBAT-GAM, B&M-ComBAT and the longitudinal-ComBAT presented in [section 2.2.1](#).

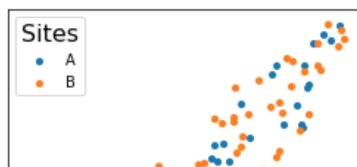
Moreover, the fact that both solutions can not be used on new sites emphasizes the need of a more suitable solution. Otherwise, it is very unlikely that any harmonization solution will be used in clinical practice.

This last point connects directly to [chapter 4](#) in which we present an original solution designed to take over various drawbacks already mentioned.



(a) Original data
AUC = 0.98

(b) After ComBAT
AUC = 0.68



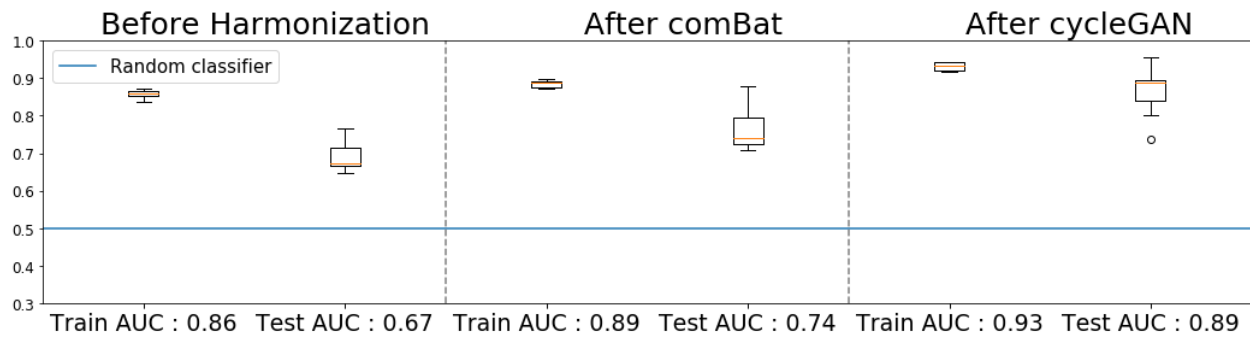


Figure 3.8 – SVM train / test AUC metric, for ASD classification on data from sites A and B (Exp5).

Chapter 4

ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization

Abstract

ImUnity is an original 2.5D deep-learning model designed for efficient and flexible MR image harmonization. A VAE-GAN network, coupled with a confusion module and an optional biological preservation module, uses multiple 2D slices taken from different anatomical locations in each subject of the training database, as well as image contrast transformations for its training. It eventually generates ‘corrected’ MR images that can be used for various multi-center population studies. Using 3 open source databases (ABIDE, OASIS and SRPBS), which contain MR images from multiple acquisition scanner types or vendors and a large range of subjects ages, we show that ImUnity: (1) outperforms state-of-the-art methods in terms of quality of images generated using traveling subjects; (2) removes sites or scanner biases while improving patients classification; (3) harmonizes data coming from new sites or scanners without the need for an additional fine-tuning and (4) allows the selection of multiple MR reconstructed images according to the desired applications. Tested here on T1-weighted images, ImUnity could be used to harmonize other types of medical images. This work was presented at SFRMBM 2021, ISMRM 2022 and OHBM 2022. It is currently under review in MEDIA.

Keywords— Brain, Deep Adversarial Network, Data harmonization, Self-supervised learning, Radiomic features

CONTENTS

4.1	Introduction	53
4.2	Materials and methods	54
4.2.1	Data	54
4.2.2	ImUnity's model	55
4.2.3	Training	57
4.2.4	Experiments	58
4.3	Results	60
4.4	Discussion	64
4.5	Conclusion	68
4.6	Compliance with ethical standards	68

4.1 Introduction

A general introduction to MRI data harmonization was presented in [section 1.2](#).

As seen in [section 2.2](#), many solutions have been proposed in the literature to tackle the issue of MR harmonization. In [chapter 3](#), we investigated the capacity of two different methods to harmonize data from two acquisition sites. This study led to interesting results, highlighting the potential of such methods to reduce site related noises while preserving biological information in order to improve further clinical studies. In addition, both solutions seemed to have complementary impacts on data. However some drawbacks make them difficult to use in clinical practice. First, they both require to be fitted every time data from a new site is added. Second, CycleGAN is intrinsically limited to two-sites harmonization. In practice, multi-centric databases gather data from much more than two sites (e.g. 11 sites in the ABIDE database).

[subsection 2.2.2](#) and [2.2.3](#) stress out the great potential of deep-learning to respond to the harmonization issue. Many approaches have been proposed in the last decade. However, there is a need for a more adapted solution, generalizable enough to be used in clinical practice. [Table 4.1](#) summarizes the main DL solutions and their intrinsic features.

	Traveling subjects	Fine-tuning for new clinical question	Fine-tuning for unseen sites	Max. number of target sites
Zhu et al. (2018) (CycleGAN)	not required	not required	required	$N = 2$
Dewey et al. (2019) (Deep-Harmony)	required	not required	required	$N = 2$
Zuo et al. (2021a) (Calamity)	not required	not required	required	$N = \text{number of training sites}$
Dinsdale et al. (2020) , Guan et al. (2021)	not required	required	not required	$N > \text{number of training sites}$
ImUnity (this study)	not required	not required	not required	$N > \text{number of training sites}$

Table 4.1 – Versatility of deep-learning harmonization models

Inspired by latest advances in harmonization generative ([Dewey et al., 2020](#); [Zuo et al., 2021a](#)) and unlearning solutions ([Dinsdale et al., 2020](#); [Guan et al., 2021](#)), we propose in this chapter a new type of harmonization method, called ImUnity. It is based on 2.5D deep-learning and extends previous techniques to offer a fast and flexible harmonization solution. ImUnity generates ‘corrected’ MR images that can then be utilized for various population imaging studies. To avoid the need for traveling subjects or multiple MR sequences in the database, our self-supervised Variational AutoEncoder (VAE-GAN) architecture uses for its training, multiple slices from the same individual and randomized image contrast transformations. It also unlearns center bias using a confusion module connected to its

bottleneck while an optional biological module can ensure that clinical features are preserved in the latent space. Once trained, this architecture should allow data coming from new sites or scanners to be harmonized without the need for fine-tuning. The architecture also allows estimates towards multiple target sites and then, users can choose multiple MR image reconstructions according to the chosen target domain (site or scanner).

To overcome the intrinsic problem of 2D generative models, i.e. the discontinuity in final outputs along the third axis, we introduce a 2.5D model that combines the outputs of 3 models, each one trained along a specific axis. This approach was first introduced to the MR harmonization field by (Dewey et al., 2019), highlighting the great potential of such an approach.

To evaluate the efficiency and flexibility of our harmonization tool, we tested the approach using 3 open source databases that contain images from multiple acquisition sites, scanner vendors or strength of magnetic fields, and a large range of patients' ages. For most of the experiments, ImUnity was trained using data from only one of the databases and then applied to the other two to evaluate generalisation of the model. Quality of the reconstructed images, capacity of removing site or scanner bias and ability to classify patients were evaluated after data harmonization.

4.2 Materials and methods

4.2.1 Data

We used three open-source databases: (1) **ABIDE**, a multi-center project led by Di Martino et al. (2014), which focuses on Autism Spectrum Disorder (ASD). It gathers more than 1,000 autistic patients and controls. For this study, we used T1-weighted scans from 11 different sites and scanners from 3 different constructors (3T scanners at 10 different sites and one 1.5T scanner at one site). Sites presenting data from a large range of ages (from 6 to 47 years, mean age = 12 years) were selected. In total, 621 T1-weighted scans (309 patients and 312 controls) were collected. (2) **OASIS** (LaMontagne et al. (2019)) gathers T1-weighted scans from healthy (N=605) and Alzheimer's Disease (AD) (N=493) adult subjects who underwent several MR sessions on 4 different scanners from the same site. We used these traveling subjects (N = 1098) to validate the ability of our model to perform multi-scanners harmonization.

(3) **SRPBS** (Tanaka et al. (2021)) is a multi-site database gathering multi-disorder subjects. We used 9 healthy adult traveling subjects to validate harmonization results between the different acquisition sites of the database (6 sites, 12 scanners from 3 different constructors). Note that SRPBS contains healthy adult brain scans while ABIDE (resp. OASIS) mainly includes healthy and pathological infant (resp. AD adult) brain scans, leading to large anatomical differences between images in the databases.

For each subject in each database, the brain was extracted using Robex (Iglesias et al., 2011) and N4Bias (Tustison et al., 2010) was used to correct for intensity inhomogeneities. MR images were first co-registered, using fsl-FLIRT (Jenkinson and Smith, 2001), to the publicly available and age specific 152-MNI templates (Sanchez et al., 2012a). Then, White-Stripe normalization (Shinohara et al., 2014) was run to align white matter (WM) peaks between all subjects (each WM peak was aligned to 0.7 after rescaling the whole image between [0:1]).

After visual inspection to detect images with ROBEX defects or other artifacts, we eventually included 545, 1072 and 81 T1-weighted scans from ABIDE, OASIS and SRPBS databases respectively.

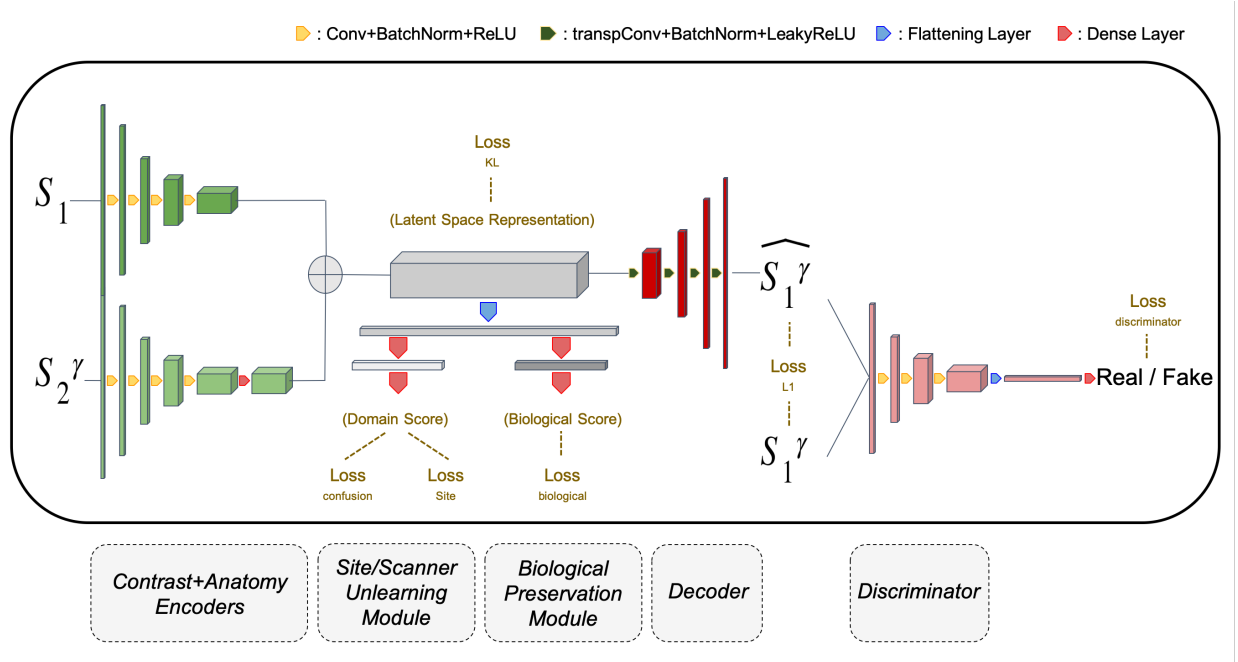


Figure 4.1 – ImUnity’s architecture. The model involves: a modified VAE generator (4.2.2), a CNN discriminator, an additional Site Unlearning module (4.2.2) and an optional Biological module (4.2.2). Here the bottleneck corresponds to the encoder’s mean and variance outputs.

4.2.2 ImUnity’s model

The architecture of our model derives from convolutional VAE-GANs and is described in Figure 4.1. We adopted adversarial settings to ensure realistic outputs using a classical CNN as a discriminator. The generator (here a VAE) learns how to represent input data into a lower dimension latent space (bottleneck). Information is then decoded to generate an output image. Inspired by Dinsdale et al. (2020), an unlearning center-bias module is connected to the bottleneck to limit the impact of site or scanner information. A biological preservation module can be inserted to maintain biological information in the latent space representation. Technical details are provided below.

Modified VAE generator

Inspired by Zuo et al. (2021a), our generator takes two 2D-structural images in the same orientation as input, randomly taken at two different locations in the 3D-MR stack of images of each subject to consider. The first (S_1) image is used by the first CNN to encode the ‘anatomical’ information using only convolutional filters to ensure preservation of spatial information. The second image (S_2) differs from S_1 because it is randomly taken in another part of the same brain and provides the initial ‘contrast’ information. At least 10 slices separate S_2 from S_1 (i.e. in our case 10 mm) ensuring that S_1 and S_2 have different anatomy

(different location in the brain) but similar contrast (same scan). S_2 contrast is modified using a gamma function (or exponential correction): $I_\gamma = range \left(\frac{I}{range} \right)^\gamma$, where I represents voxels intensities and ‘range’ is the difference between maximum and minimum intensity values. The ‘gamma’ parameter γ is sampled uniformly between 0.5 and 1.5 for each new input 2d slice. This modified S_2^γ slice is used as input to a second CNN to encode the ‘contrast information’ followed by a dense layer to reduce spatial information. An example of different gamma transformations applied to MR brain scans from the same subject is given in [Figure 4.2](#). Once encoded, the two independent representations of S_1 and S_2^γ are concatenated to give a latent space representation which is decoded to create the output \hat{S}_1^γ using transposed convolutional filters. Eventually, this output is compared to the reference gamma modified slice S_1^γ . Note that this generator is trained in a self-supervised fashion as output labels are generated during the training phase and this training can be done on any MR dataset. It does not require additional information such as scanner, center or biological information.

Site/Scanner-bias unlearning module

To ensure the task of “removing site or scanner bias”, a module is directly connected to the encoders’ outputs (latent space representation of inputs). The module can be seen as a domain (site or scanner) discriminator and is trained independently from the encoder to predict the scan’s origin based on the latent space representation. On the other hand, the encoder is trained in an adversarial fashion. A confusion loss is used to unlearn domain information. This principle has been introduced in the field of domain adaptation by [Ganin et al. \(2016\)](#) and has been adapted to medical imaging studies by [Dinsdale et al. \(2020\)](#). Originally, the module was incorporated directly in the model to unlearn datasets bias and to improve predictions. Here, it is used in the bottleneck as a “datasets bias filter”, forcing the encoder to learn a domain-invariant data representation. Note that the architecture of ImUnity differs from that of [Dinsdale et al. \(2020\)](#), and so does the position of the bottleneck. Overall, the generator learns a shared latent space that encodes all information needed to generate harmonized scans. We also chose to avoid skip-connections in our network to ensure that site/scanner related information (present in input data) does not flow directly through these connections (as it would be the case in a UNet ([Ronneberger et al., 2015](#)) for example). The loss function for the site/scanner unlearning module is:

$$l_{site}(P, Y) = -\frac{1}{N} \sum_{i=1}^N \sum_{s=1}^S \mathbb{1}(y_i = s) \log(p_i^s) \quad (4.1)$$

While the confusion loss used in the encoders’ training is :

$$l_{confusion}(P) = -\frac{1}{N} \sum_{i=1}^N \sum_{s=1}^S \frac{\log(p_i^s)}{S} \quad (4.2)$$

Here $P = [p_1; \dots; p_S]$ is the softmax output from the module, corresponding to the probability to belong to different sites (1, ..., S) , Y is the ground truth site affiliation vector, and N is the sample size.

Biological preservation module

An optional module ensures the “preservation” of biological information. It acts as a classifier of available biological information. For instance, features such as age or the presence of diseases can be introduced. Contrary to the unlearning module, the encoder is trained to minimize its loss function. This module is not mandatory, and a fully self-supervised learning model can be adopted when it is turned off. Then, the loss function of the biological preservation module for our particular application using the ABIDE database is:

$$l_{biological}(P, Y) = -\frac{1}{N} \sum_{i=1}^N \sum_{f \in features} \sum_{i=1}^N y_i^f \log(p_i^f) + (1 - y_i^f) \log(1 - p_i^f) \quad (4.3)$$

Here P represents module predictions for biological features of interest, N is the sample size, and Y is the ground truth vector. Note that in this study, the binary cross entropy formulation was used for the loss function because only two features (age and patient status, i.e. ASD) were considered.

A 2.5D solution

The presented architecture is in 2D. As presented in [Figure 4.4-right panel](#), one can see that a 2D model generates high-quality images along the training axis. However, its final 3D reconstruction suffers from artifacts along the two other axes. This is why in this study, we propose to use the above model in a 2.5D way. It consists of using three 2D models along each axis and to combine their results in order to have sharper output. We fuse predictions using the median value, an approach less sensitive to outliers than the mean value used in [Dewey et al. \(2019\)](#). Using a 2.5D approach generates higher quality images while keeping a number of parameters reasonable ($\sim 5.10E6$ for each model) without requiring more training datasets. The introduction of a 3D architecture would highly increase the number of parameters to estimate ($\sim 30.10E6$), therefore requiring more training datasets.

4.2.3 Training

For each model, training involves several independent steps, due to the adversarial context and the use of the additional modules.

- Training the discriminator consists in minimizing the binary cross-entropy $l_{discriminator}$ between its predictions and the labels corresponding to the nature of the inputs (real or fake). Adversarially, the generator learns how to maximize this loss function, forcing the generation of realistic outputs.
- Training the site/scanner unlearning module consists in minimizing the categorical cross-entropy ([Equation 4.1](#)) between its predictions and the site-affiliation labels. Adversarially, the generator is trained to minimize the confusion loss ([Equation 4.2](#)). It forces a site and scanner invariant representation of the dataset in the latent space, leading to uniform outputs of the unlearning module.

- In our ABIDE experiment, training the biological preservation module consisted in minimizing binary cross-entropy losses associated with each biological feature taken into account (here sex and patient status). Unlike the previous module, the loss $l_{biological}$ was directly integrated into the generator. This ensures the conservation of biological features in the latent space.
- In addition to previous loss functions involved in training the generator, a l_1 loss function is used to ensure a good mapping between input $(S_1; S_2)$ and the generated output \hat{S}_1^γ ($l_1 = mean(|\hat{S}_1^\gamma - S_1^\gamma|)$). Moreover, the use of the Kullback-Leibler divergence l_{KL} (see Equation 4.5) between feature distributions and a Gaussian distribution ensures a dense data representation in latent space. Therefore, the global generator’s loss function to minimize is:

$$l_{generator} = -\lambda_1 l_{discriminator} + \lambda_2 l_{confusion} + \lambda_3 l_{biological} + \lambda_4 l_1 + \lambda_5 l_{KL} \quad (4.4)$$

$$l_{KL}(P, Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (4.5)$$

Here, λ factors control the relative contribution of each loss. In our study, we used: $\lambda_1 = 1$; $\lambda_2 = 1$; $\lambda_3 = 1$; $\lambda_4 = 100$; $\lambda_5 = 10^{-3}$ found empirically and with P and Q, two discrete distributions. Note that because the generator searches to fool the site-classification module by forcing an uniform prediction, we integrate the confusion loss (Equation 4.2) and not the module’s loss (Equation 4.1) in Equation 4.4.

Note that for all the following experiments, ImUnity-2.5d was used. The code of ImUnity is in open source access here: https://github.com/nifm-gin/dl_generic.

4.2.4 Experiments

The datasets extracted from the three databases were used to evaluate different aspects of our model. The impact on image quality in multi-site or multi-scanner harmonization was assessed using data from traveling subjects (ground truth) from the OASIS and SRPBS datasets. Ability to remove site information was evaluated using the ABIDE dataset. Finally, the benefits of harmonization between data provider centers were assessed using autism disorder prediction in children from the ABIDE dataset. To demonstrate the flexibility of ImUnity, all experiments were performed with the same model trained on data coming from the ABIDE database, unless specified. OASIS and SRPBS were used for the validation parts only. Each model was trained on 2D slices with at least 1% of brain tissue voxels. Training was run on a Nvidia GeForce 2080 RTX for 300 epochs using a learning rate of 10^{-4} and Adam optimizer.

Experiment 1 : Harmonization on traveling subjects (OASIS+SRPBS)

We first evaluated the ability of our model to transform images from one domain (site or scanner) to their equivalent in another domain. As SRPBS and OASIS databases contain

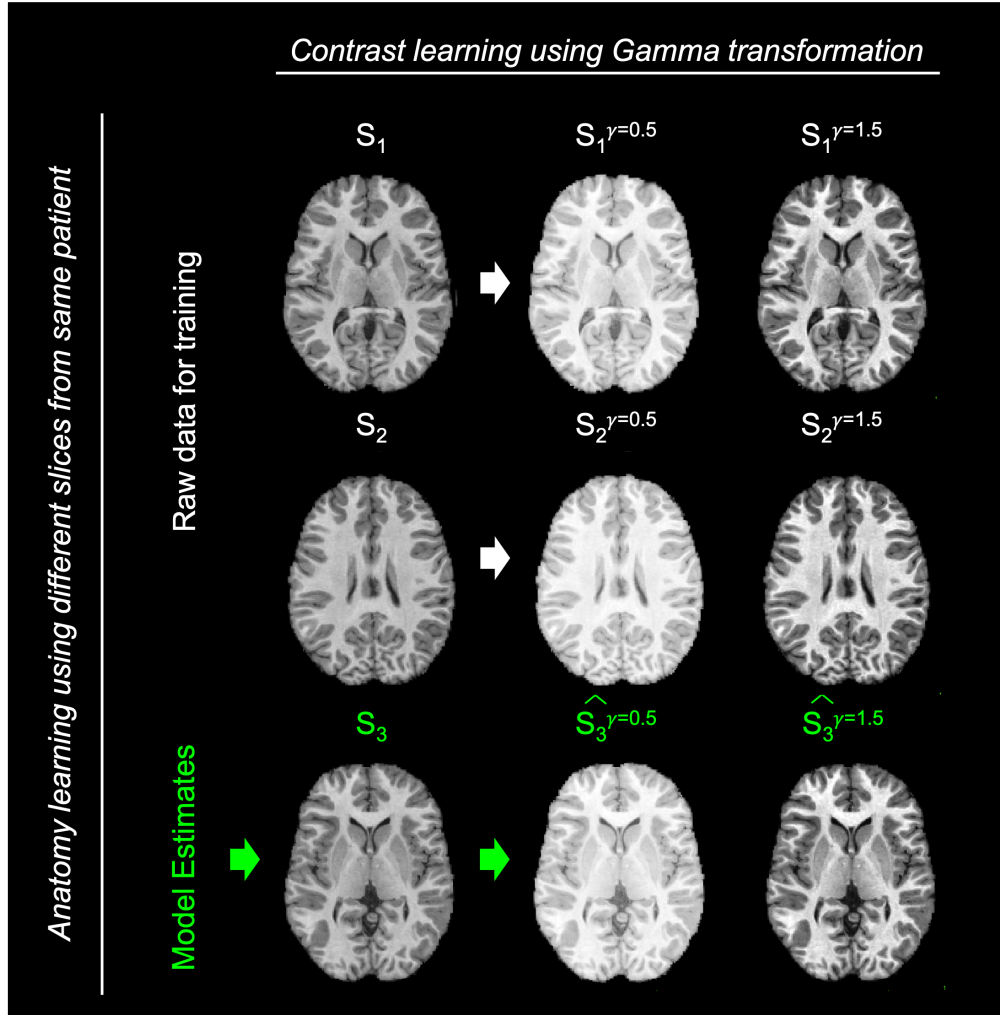


Figure 4.2 – Inputs and outputs of the model: Different slices from the same patient are used to encode the anatomical information. Gamma transformations are used to encode the contrast information. Rows represent different anatomical slices (S_1 , S_2 and S_3) taken from the same subject. The first two rows present Gamma transformations used to train the model (S_1^γ and S_2^γ). The last row shows model outputs (\hat{S}_3^γ) for estimated Gamma transformations for the slice S_3 not present in the training set. From left to right columns: original slices, Gamma modified slices with parameter 0.5, Gamma modified slices with parameter 1.5. Three 2D models along each direction are combined for the 2.5D approach.

traveling subjects, ground truth was available to assess ImUnity performance. In practice, one domain (acquisition site or scanner) was first selected as the reference for every subject. Individual scans were co-registered to their equivalent in the reference domain (to avoid variations between acquisitions due to movement). Then, all the images were transformed by the model into the reference domain. During this step, the slices to be harmonized (anatomy) were fed to the model along with the corresponding computed contrast slice from the reference domain. Finally, results obtained after transformation were compared to the ground truth, i.e. images acquired in the reference domain (traveling subject). Visual verification, image intensities histograms, and the Structural Similarity Index Metric (SSIM, Wang et al. (2003)) were used to assess image likeness. Paired t-tests were used for statistical significance.

Furthermore, the same model, trained on ABIDE data, was used for every site / scan of the other two databases to evaluate ImUnity’s ability to generalize to sites never seen before. This experiment also evaluates ImUnity’s versatility, either for the source domain or for the target domain (last two columns of [Table 4.1](#)). Additionally, we also trained and tested ImUnity on different sites combinations to evaluate the impact of the sample size and the population type.

Experiments 2 : Harmonization’s effects on sites classification (ABIDE)

The second experiment evaluated the ability to detect the origin of data before and after harmonization. Harmonized data were obtained using a 5-fold cross-validation procedure on the ABIDE data. As no ground truth was available for this experiment, we considered harmonization impacts on classification algorithms. Standard Support Vector Machine (SVM) with a radial basis function kernel was used to classify the ABIDE data. The classifier worked on all radiomic features (N=101) extracted using the pyradiomics python API ([van Griethuyzen et al., 2017](#)). These features aim to represent different aspects of MRI images, such as shape, contrast, or texture, and are known to be sensitive to site effects ([Orlhac et al., 2019](#)). The most ‘correlated features’ with sites affiliations before harmonization were selected for classification using Pearson tests (ran independently for each feature) using 10^{-3} as p-value threshold (30 features in total). Accuracy and Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve were used to evaluate the specificity and sensitivity of the site classifier.

Experiments 3 : Harmonization’s effects on autism syndrome disorder prediction (ABIDE)

Similarly to Experiment 2, Experiment 3 evaluated the ability of our classifier to detect patients with ASD from the ABIDE database, before and after harmonization. Here, results were obtained following a 10-fold cross-validation procedure. The same trained model was used for different numbers of sites (and different combinations of sites) included in the ABIDE database.

4.3 Results

Experiment 1: [Figure 4.3-A](#) shows the results obtained for one traveling subject from the SRPBS database ([section 4.2.4](#)). Images are shown for one acquisition at site (A) before harmonization ([Figure 4.3-A](#), left), corrected by ImUnity to fit with acquisition at site B ([Figure 4.3-A](#) middle) and the corresponding ground truth acquired at site B ([Figure 4.3-A](#) right). One can notice the difference in image contrast between the 2 sites, highlighting the need for image harmonization, as well as the visual similarity between the harmonized image and the ground truth. It is interesting to observe that the anatomical structures of the input contrast reference are not propagated through the model, which explains small anatomical differences (e.g. superior sagittal sinus) between the model estimates and the ground truth. It is also worth noting that although each model was trained on 2D slices, combining outputs by taking the median for each voxel gives a final 3D reconstruction of the estimates of high quality in each orientation. [Figure 4.4](#) highlights the positive impact of the 2D models fusion. For some subjects, 2D models present artifacts along the third axis that disappear after the

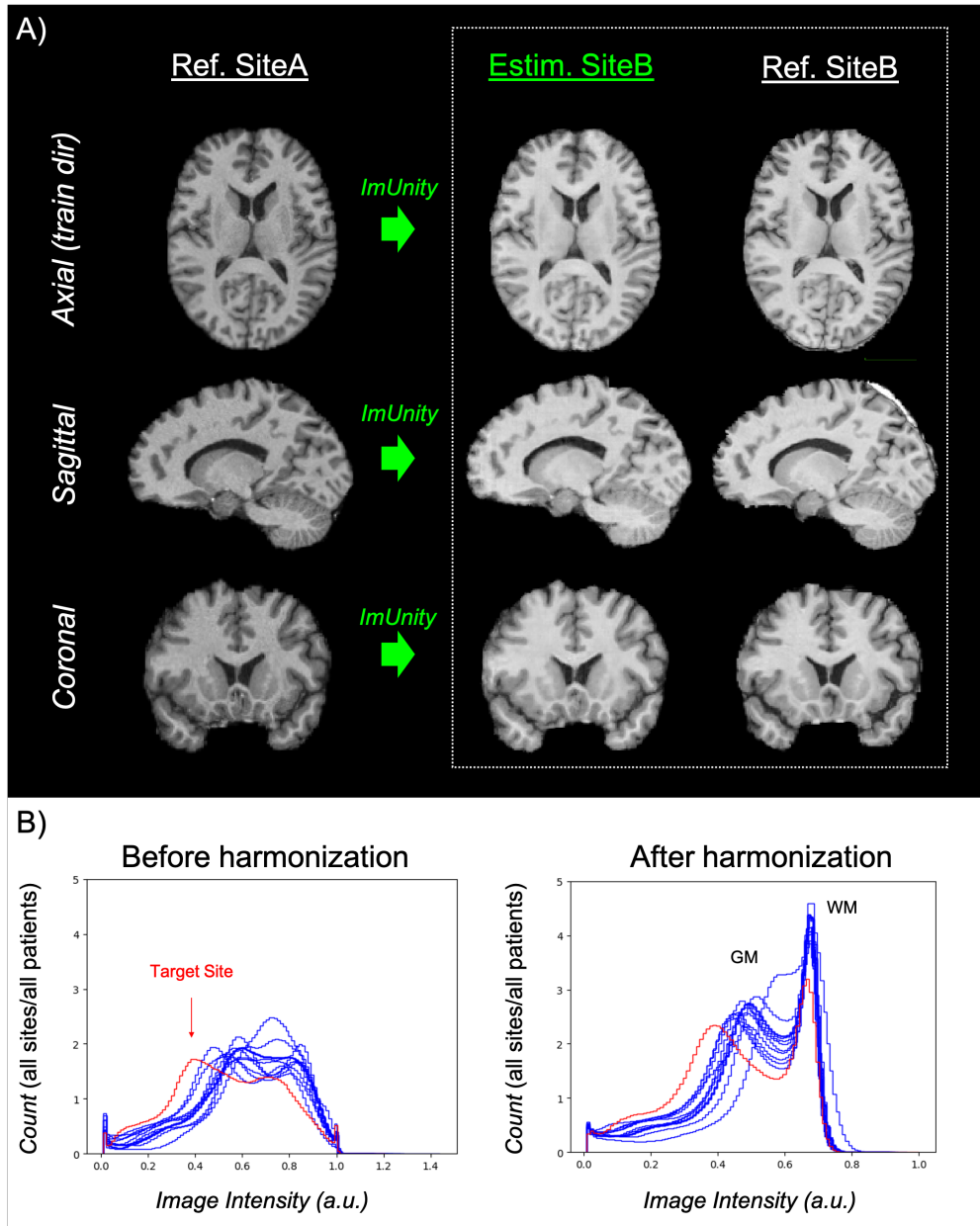


Figure 4.3 – Harmonization results on traveling subjects from the SRPBS database. A) Left: 3D images from one patient (axial, sagittal, coronal views) acquired in site A before harmonization, Middle: ImUnity’s harmonization to fit with acquisition at site B, Right: image acquired at site B (ground truth). B) Images intensity distributions (all patients) before (Left) and after ImUnity’s harmonization (Right). The red histogram corresponds to the site taken as reference for the harmonization process (target site). GM = gray matter; WM = white matter

median 2.5D combination. Figure 4.3-B shows the effects of ImUnity’s harmonization on image intensity distributions for all selected subjects from the SRPBS database. The model was used to harmonize every image to a target site (indicated in red). An alignment of histograms is clearly observed after harmonization, with both gray and white matter peaks shifted. Changes in intensity distribution of the site of reference are due to pre-processing

(see details in Figure 4.6, top row). The images obtained after ImUnity’s harmonization of the OASIS datasets are also provided in Figure 4.5.

Task	Multi-scanner harmonization.		Multi-site harmonization
Dataset	OASIS scanner F → scanner E	OASIS all scanners → scanner E	SRPBS all sites → UTO site
Raw data	0.871 ± 0.045	0.845 ± 0.059	0.853 ± 0.021
Zhu et al. (2018) CycleGAN*	0.873 ± 0.046	-	-
Zuo et al. (2021a) Calamity*	0.884 ± 0.046	-	-
ImUnity [#]	0.918 ± 0.071 **	0.920 ± 0.067 **	0.882 ± 0.060 **
ImUnity*	0.951 ± 0.013 **	0.942 ± 0.011 **	0.901 ± 0.035 **
ImUnity ⁺	0.832 ± 0.067	0.835 ± 0.063	0.883 ± 0.059

Table 4.2 – SSIM in traveling subjects for multi-scanner (healthy subjects from the OASIS database) and multi-site (SRPBS database) harmonization. Results are compared to the literature when available.

*: Model trained on OASIS database (n=1072); #: Model trained on ABIDE database (n=545); +: Model trained on SRPBS database (n=81) **: Significant improvement ($p < 10^{-5}$)

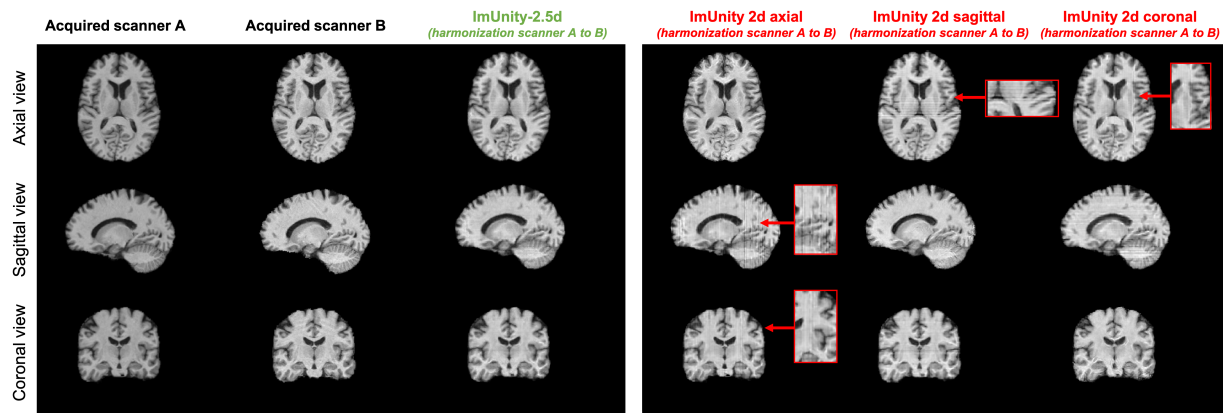


Figure 4.4 – Impact of the 2.5d median fusion approach (section 4.2) on one traveling subject from the OASIS database, spotting out some discontinuities when using only one 2d ImUnity model.

Quantitative results obtained with the SSIM metric in all traveling subjects are summarized in Table 4.2. Both multi-site (SRPBS) and multi-scanner (OASIS) experiments are

shown. For the latter, results from the literature are also given for reference. It can be seen that when trained on the ABIDE database, ImUnity significantly increases the structural similarity in all cases and provides better performances compared to other deep learning approaches. Moreover, results from multi-scanner harmonization show that ImUnity performs well independently of the chosen reference domain. The last 2 rows in Table 4.2 present results obtained after training ImUnity on OASIS (to better match literature protocols) and SRPBS (to highlight sample size impact on the model’s generalization) data. These models were used to harmonize OASIS as well as SRPBS data. Note here that multi-scanner harmonization results were obtained on healthy subjects only to match literature results. Similar results were obtained when including AD subjects (not shown).

In the same way, Table 4.3 presents the additional quantitative results for multi-scanner harmonization. In this table we can observe a significant improvement (in our case, a decrease) of MAE and MSE metrics after ImUnity data harmonization. Confirming the positive effect of the proposed solution.

	OASIS scanner F \rightarrow scanner E		OASIS all scanners \rightarrow scanner E	
	MAE	MSE	MAE	MSE
Before ImUnity	0.0224 \pm 0.0108	0.0040 \pm 0.0045	0.0232 \pm 0.0108	0.0042 \pm 0.0044
After ImUnity	0.0139 \pm 0.0009	0.0012 \pm 0.0002	0.0142 \pm 0.0018	0.0013 \pm 0.0003

Table 4.3 – MAE and MSE in traveling subjects for multi-scanner (OASIS database) harmonization. Here the model used was trained on the ABIDE database only (n=545). Results in bold present a significant ($p < 10^{-5}$) improvement after ImUnity.

Experiment 2: 4.7(a) shows ImUnity’s harmonization effects on site classification on the ABIDE datasets (section 4.2.4) using tSNE (Maaten and Hinton, 2008), a dimension reduction algorithm, on radiomic features. Before harmonization, the presence of site clusters is clear. Once the data are harmonized using ImUnity, the points are shuffled and the accuracy of the SVM site prediction decreases from 0.74 to 0.37 (before and after harmonization respectively). This confirms the removal of site bias by ImUnity as the classifier is no longer able to correctly separate the sites. Note that 4.7(a) also shows that small clusters remain after harmonization, which could be explained by remaining site or scanner bias or by difference in demographic (age or sex) or biological (pathology) features between the respective groups of subjects. Additional results on the influence of the pre-processing step on site classification are provided in Figure 4.6.

Experiment 3: 4.7(b) shows the capacity of our model to improve ASD prediction from the ABIDE datasets. Here, we used the same trained models to test the influence of different numbers of sites included in the database (from 2 to 11) as well as different combinations of those sites (for example 55 combinations of 2 sites taken among the 11 sites available). In every case, we observed a clear improvement of classification of autistic patients after harmonization as shown by increases in AUC provided by the SVM classifier. We show the results obtained with the best combination of sites as well as average and standard deviation of AUC with all combinations of sites. The pre-processing also has a positive impact on the prediction as shown in Figure 4.6-bottom row.

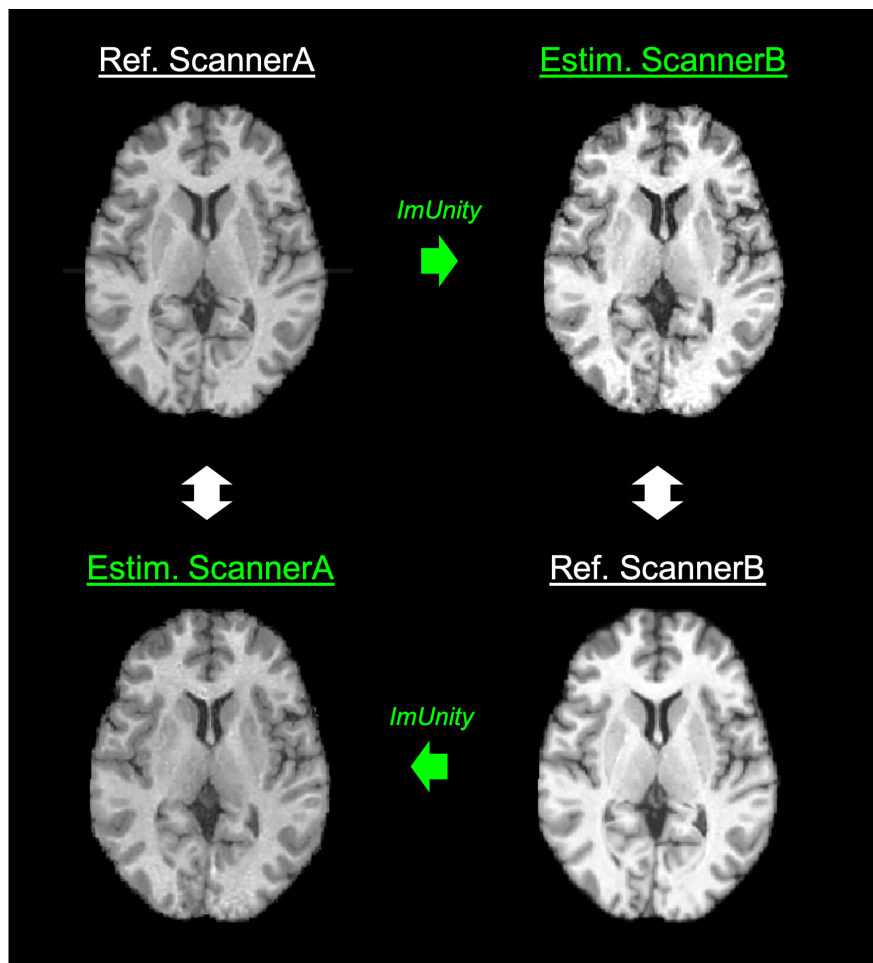


Figure 4.5 – Multi-scanner harmonization (section 4.2.4) results between 2 scanners for the same subject extracted from the OASIS database. ImUnity’s model was trained on datasets extracted from the ABIDE database. Top row: Slice acquired at site A (left) and corresponding harmonized image (right) matching acquisitions at site B. Bottom row: Slice acquired at site B (right) and corresponding harmonized image (left) matching acquisitions at site A. Left (resp. right) column allows to visually compare the ground truth and the estimated image for site A (resp. for site B).

4.4 Discussion

We have presented ImUnity, an original 2.5D harmonization tool for multi-center MRI databases. ImUnity shows high performances in terms of quality of the generated harmonized images, as well as clear removal of the idiosyncratic bias attached to site-dependent image acquisition conditions. Moreover, the performed experiments clearly demonstrate ImUnity’s versatility. By training ImUnity’s model on datasets extracted from one database (here ABIDE) and looking at images harmonized from traveling subjects provided by two different databases (here OASIS and SRPBS), we show that ImUnity does not require new training phase to generalize to unseen sites or scanners (see Figure 4.3). The performances were maintained independently of the site selected as reference (see Table 4.2). While the model was trained on ABIDE data only, it provided better results than the state-of-the-art methods in terms of image quality (+4%, see Table 4.2).

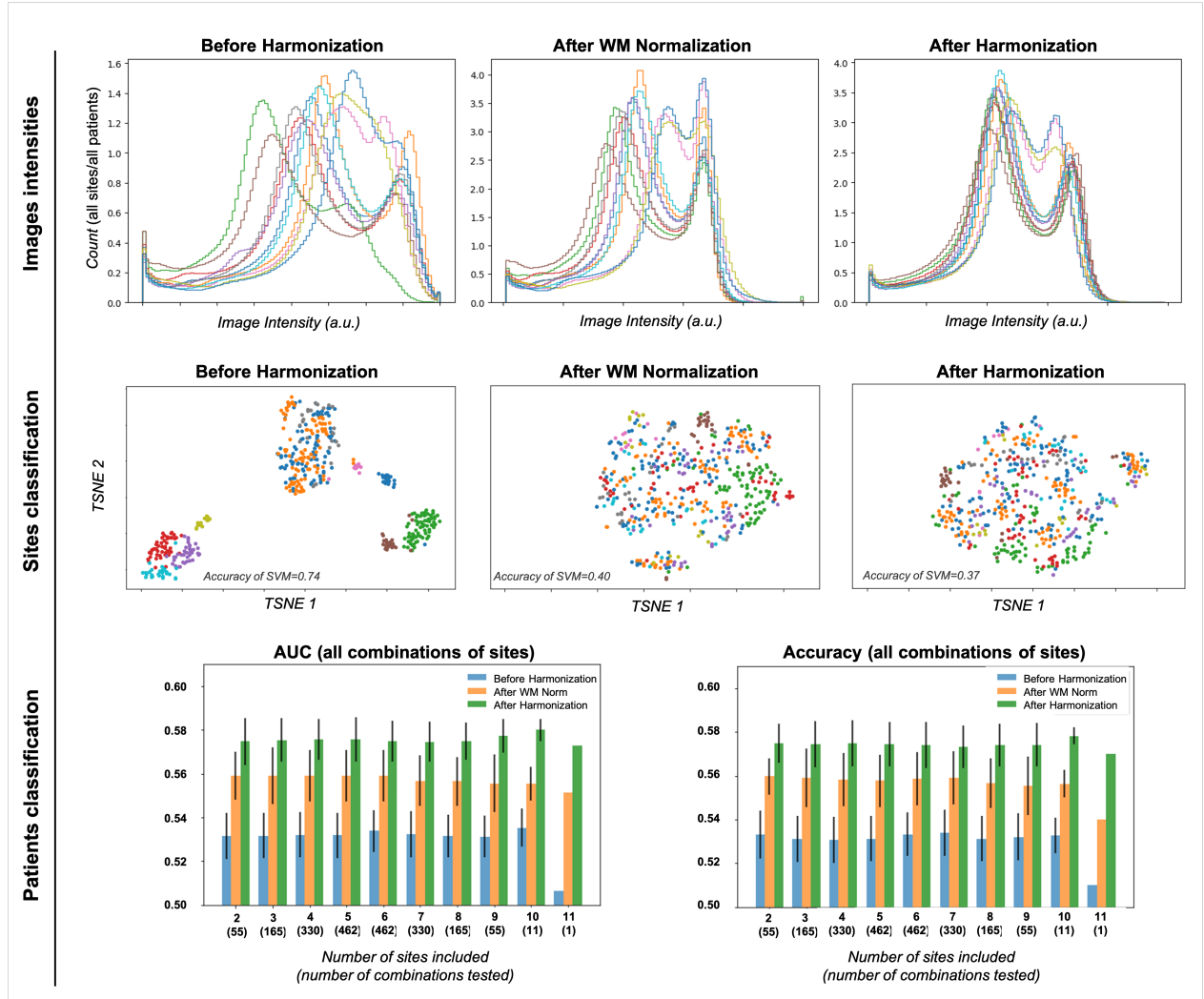


Figure 4.6 – Impact of pre-processing steps (N4Biais, White-Stripe normalization) on our different experiments. From top to bottom row: impact on images intensity, impact on sites classification (4.2.4), impact on patients classification (4.2.4)

The last two rows of Table 4.2 present SSIM metrics obtained when training ImUnity on the other 2 datasets (OASIS and SRPBS). As no biological features were available in these databases, the biological module was disabled and the model was trained in a self-supervised way. First, we noted additional improvements for scanner harmonization when the model was trained and applied on the same database (here OASIS, +2.5%). Second, the score obtained for multi-site harmonization (SRPBS) was the highest when trained on OASIS ($N = 1098$) data (with a slightly better score than with the other databases). It is interesting to observe the impact on these scores of the dataset size, the number of site/scanner involved in the training, the use of the biological module and the anatomical differences between datasets (ABIDE mainly contains children data while OASIS and SRPBS focuses on adults). While OASIS results suggest a better generalization on unseen data because more training data were available, ABIDE results suggest that anatomical differences could be compensated by a large training dataset presenting more site/scanner variability than OASIS (11 sites for ABIDE vs. 4 sites for OASIS). On the other hand, results from multi-scanners harmonization

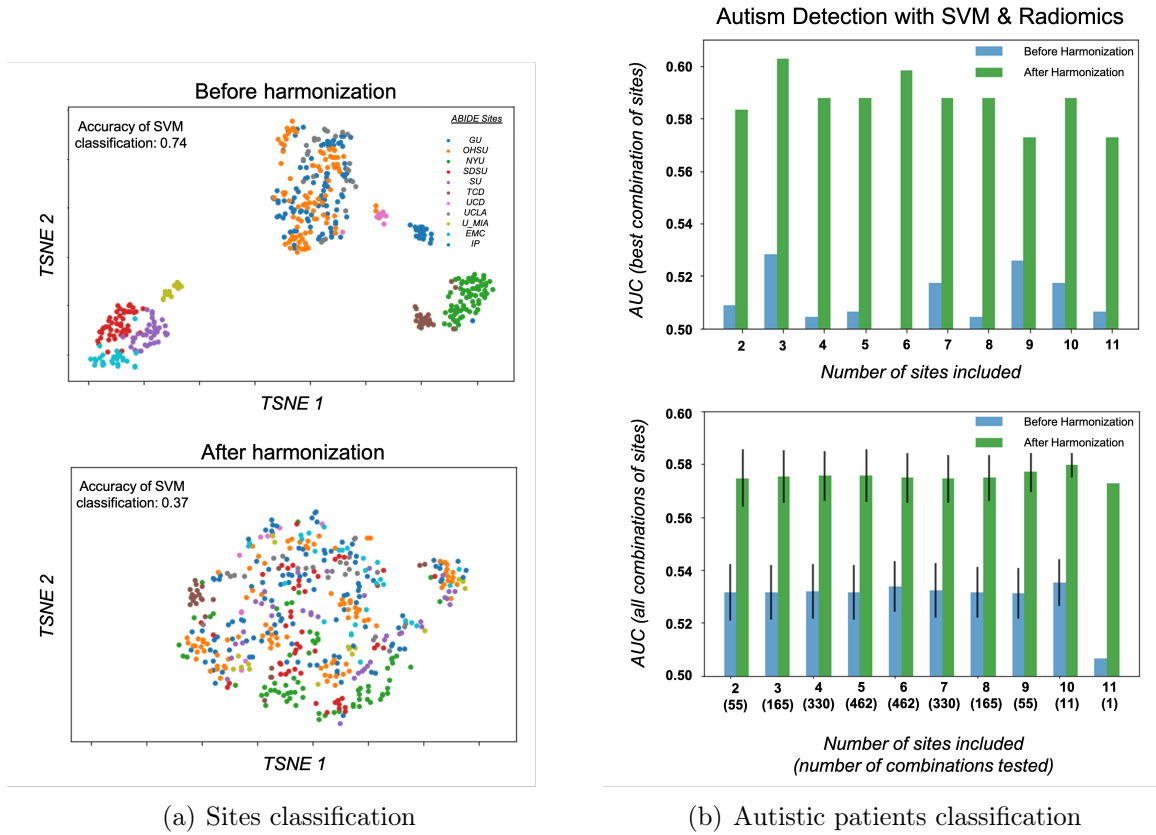


Figure 4.7 – **A)** Harmonization effects on ABIDE sites classification. 2D scans representation of ABIDE database using tSNE reduction algorithm are presented before and after harmonization. Colors correspond to different sites. **B)** Harmonization effects on ABIDE patients classification. AUC metrics for classification of patients with autism spectrum disorder (ASD) using SVM and extracted radiomic features (Experience 3, 4.2.4), are shown for different numbers of sites included in the database (from 2 to 11 sites). Top row: Results obtained from the best (largest change in AUC before and after harmonization) combination of sites. Bottom row : Average and standard deviation of AUC estimates for all combinations of sites. The same trained model and harmonized data were used for different site combinations.

depict the difficulty of the model trained on SRPBS data to generalize its training to the OASIS data. This indicates an over-fitting effect in this situation, as there was not enough training data (here $N=81$ distributed over 9 sites). This suggests that ImUnity may not be adapted to small sample size scenarios, which provides useful information for understanding why the model is less effective in some contexts. Note that data augmentation could have been performed to improve the results in this experiment.

To better estimate the impact of the biological module, we have also run Experiment 3 (section 4.2.4) with this option disabled. We found (see Figure 4.8) that the biological module has a positive impact on the results with a contribution representing about 20 percent of the total harmonization effect and suggesting that additional input features along the image could lead to better harmonized outputs. Additionally, in Experience 1, we tested different combinations of training and testing data. In the situation: 'training = OASIS' and 'testing

= SRPBS’, no biological features were available, thus the biological module was not used. Our solution still showed good generalization on the testing dataset. Based on this set of results, we may assume that the addition of some biological information to the model would also lead to better results on traveling subjects. Because ImUnity is designed to reconstruct images and to create a new harmonized database, it does not need new training for new clinical or biological questions. Beyond classification, new clinical data investigation should be conducted with ABIDE (or other multi-center clinical databases) to have better understanding on the impact of our method on clinical research studies.

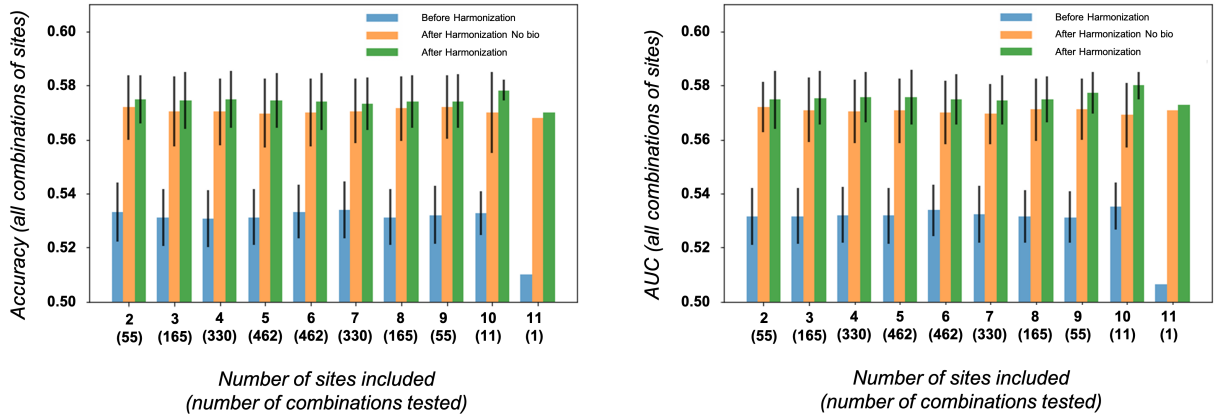


Figure 4.8 – Impact of the biological module on experiment 3 (4.2.4). Patient classification accuracy and AUC metrics are presented for data before harmonization (blue), after ImUnity without biological module (orange) and after ImUnity (green).

Like the majority of deep networks used for medical image analysis, the MR images used as inputs of our network were first pre-processed for intensity normalization, co-registration or brain extraction. Usually, the impact of these transformations is not examined in harmonization studies. Figure 4.6 highlights the fact that these steps are already able to remove some of the sites and scanners biases with positive impacts on intensity distributions across sites or patients classification. Intrinsically, the use of White-Stripe normalization (Shinohara et al., 2014) forces the alignment of intensity distributions. Yet, a perfect alignment is not the ultimate goal of harmonization as we also seek to preserve informative biological variations which should persist independently across sites. Eventually, we observe that the best results are obtained for all experiments after the whole ImUnity process, with better aligned intensity distributions, removal of persistent datasets’ noises, and most importantly improvement in patients classification results. On the contrary, other experiments (not reported) also showed that the VAE-GAN network alone performed poorer when the pre-processing steps were omitted, suggesting that these steps are needed to simplify the training process and improve generalization of the results.

As harmonization is a direct application of domain adaptation, we could investigate different architectures like Choi et al. (2018); Huang and Belongie (2017) and include Adaptive Instance Normalization instead of classical batch normalization layers. Then, very promising results were reported with the use of the StarGAN model for MR harmonization (Bashyam et al., 2021). Finally the inclusion of a cycle consistency loss as presented in (Liu et al., 2021) could enhance our contrast encoder and force a good contrast-style representation in the latent space.

In our study, we only report results with anatomical T1-weighted images. We show that a single type of sequence, combined with computed image transformations with the Gamma function, are sufficient to learn contrast mapping. This greatly facilitates the use of our model because of few data requirements (the origin of each scan is the only pre-required information) and the possibility of self-supervised training. Yet, we believe that this approach is not only dedicated to T1 contrast harmonization and can easily be generalized to any MRI sequences. Presently, the model needs to be fine-tuned in order to harmonize a new medical imaging type. It could however be interesting to investigate its capacity to learn how to harmonize multiple sequences at once. This could be done by mixing sequence types in our training dataset and ensuring the conservation of this information by adding a new conservation module in the bottleneck. It could also be interesting to add other types of artificial contrast transformations (Schettini et al., 2010) for our training in order to account for other types of sites or sequence biases. Similarly we could also include global geometrical distortions to account for biases impacting images geometry. Finally, although using a 2.5D approach is time efficient, it inherently limits the quality of generated images compared to a fully 3D approach. However, this would induce more parameters ($\sim 32.10E6$) to estimate and therefore would require more computational power and the availability of larger training datasets.

4.5 Conclusion

We presented ImUnity, an original and effective tool dedicated to MRI harmonization. Our proposed 2.5D model derives from the VAE-GAN architecture. It ensures realistic outputs and allows removal of idiosyncratic datasets bias and the preservation of biological information. Our results show that the method reaches state-of-the-art performance in terms of image quality on traveling patients of the OASIS and SRPBS databases and improves autistic patients classification from the ABIDE database. The proposed 2.5D model is versatile, requiring only one type of MR sequence without the need of matching subjects, can be generalized to sites unseen during the training phase and can be used to harmonize MR images to different reference domains without a new training phase.

Perspectives of enhancement still remain. The introduction of more complex contrast variations during the training phase and the use of domain adaptation techniques could benefit our proposed solution, especially for multiple MR sequences harmonization. Even if the proposed solution has been validated in a clinical experiment and has proven to be effective, it still needs to be tested more generally. Several harmonization studies (Beer et al., 2020; Fortin et al., 2018) have studied the impact of harmonization on brain development, on which we have strong results from large mono-centric studies as presented in subsection 2.1.1. The next chapter (chapter 5) focuses on this point, evaluating the impact of ImUnity on brain development and comparing it to ComBAT.

4.6 Compliance with ethical standards

This research study was conducted retrospectively using human subject data made available by the following open sources: **ABIDE**, **OASIS**, **SRPBS**. Ethical approval was not required as confirmed by the license attached with the data.

Chapter 5

Harmonization impact on brain structure volume and thickness evolution with age on ASD and control subjects

Abstract

In this chapter, we propose to investigate the impact of ImUnity’s harmonization on brain volumetric and thickness analysis. As seen in [chapter 4](#), promising results have been obtained using multiple databases and under different experimental conditions such as increasing structural similarity index between traveling subjects or improving autism patients classification. In the present chapter, we evaluate ImUnity for its capacity to improve the study of brain volume and thickness developmental trajectories with age. As in [chapter 3](#), we also chose to compare ImUnity to ComBAT in order to push further the validation process of our method. We gathered data from 271 healthy and 253 ASD children from 5 to 25 years old, acquired at 11 sites provided by the ABIDE database. We used Freesurfer to investigate the effect of harmonization on the evolution with age of 28 brain regions of interest. Comparing the results to those obtained in the literature, we could demonstrate the need for image harmonization and the positive effect of ImUnity in most regions of interest. Indeed, ImUnity corrects the volume evolution trajectories (linear or quadratic) to bring them closer to the literature. Comparing ImUnity to ComBAT (used here to directly harmonize T1w images), we showed that ImUnity provided better estimates of volume trajectories (e.g. putamen, thalamus or accumbens nucleus) in both control and ASD subjects. These results were confirmed with a statistical approach based on a mixed linear model effects model. Additionally, we could not observe such positive impacts of ImUnity nor ComBAT on thickness evolution for the regions of interest investigated.

This work was conducted with Constance Sohler, a engineer student who realized her ‘end of study project’ working on this topic under my supervision.

Keywords— Brain, Data harmonization, Volumetric, ASD, Thickness, Radiomic features

CONTENTS

5.1	Introduction	71
5.2	Materials and methods	71
5.2.1	Data	71
5.2.2	ImUnity	72
5.2.3	ComBAT	72
5.2.4	Data preprocessing	72
5.2.5	Experiments	73
5.3	Results	77
5.3.1	Volumetric analysis on healthy subjects	77
5.3.2	Thickness analysis on healthy subjects	81
5.3.3	Impact on ASD patients	83
5.3.4	Mixed Linear Model analysis	84
5.3.5	Freesurfer group analysis	86
5.4	Discussion	87

5.1 Introduction

Refer to [section 1.2](#) for a general introduction to MR-harmonization.

Several harmonization methods have been proposed in the last few years in response to the creation of large open access databases ([Fortin et al., 2016](#); [Shinohara et al., 2014](#)). A popular harmonization approach is the statistical ComBined Association Test (or ComBAT), already introduced and presented in [subsection 3.2.2](#). Nonlinear harmonization methods that use image generation networks and domain adaptation techniques have also been recently proposed ([Dewey et al., 2019](#); [Zuo et al., 2021b](#)). Amongst them, ImUnity, presented in [chapter 4](#) is an original deep learning model designed to avoid some practical harmonization difficulties. ImUnity does not require data from traveling subjects for its training, its architecture allows users to harmonize images contrast into an arbitrary chosen contrast domain and, once trained, the model can easily be used to harmonize images that were never seen before, without the need for fine tuning. ImUnity has already been tested using traveling subjects from two databases ([LaMontagne et al., 2019](#); [Tanaka et al., 2021](#)) and out-reached state of the art metrics between the multiple acquisitions although it was trained on a different database. Moreover, it was able to remove site or scanner biases of the ABIDE database ([Di Martino et al., 2014](#)) while improving classification of patients with Autism Spectrum Disorder (ASD).

In the present chapter, we assess the quality of harmonized reconstructed images by conducting a brain development study on subjects from 5 to 25 years old, taken from the ABIDE database. Previously, in [chapter 3](#) and [chapter 4](#), we observed an important need for harmonization when running a multi-centric analysis on the ABIDE database. These unwanted variations are likely to impact the apparent biological brain development, altering in consequence regions of interest (ROIs) evolution tendencies with age. Here, data coming from 11 sites were pooled and analyzed before and after images harmonization. For every subject (of different ages), volumes and thickness of cortical and subcortical ROI were extracted from original and harmonized MR images. Volume and thickness trends, before and after data harmonization, were compared to literature trends previously obtained from large monocentric studies. For comparison purposes, we also ran ComBAT harmonization directly on the images (see [section 2.2.1](#)).

5.2 Materials and methods

5.2.1 Data

As for [chapter 4](#), the data used in this study comes from three open source databases. That is to say:

- **OASIS** [LaMontagne et al. \(2019\)](#) gathers T1w scans from adult subjects who underwent several MR sessions on 4 different scanners from the same site;
- **SRPBS** [Tanaka et al. \(2021\)](#) is a multi-site database containing nine healthy adult traveling subjects acquired at nine different centers, making a total of 81 T1w scans;
- **ABIDE** , a multi-center project [Di Martino et al. \(2014\)](#), which focuses on Autism Spectrum Disorder (ASD). It gathers more than 1,000 ASD infants and healthy controls. From ABIDE, we selected T1w healthy and ASD scans from 11 different sites and scanners from 3 different constructors (3T scanners at 10 different sites and one 1.5T

scanner at one site). We selected only subjects between 5 and 25 years old, because beyond that too few subjects were represented. In total, 524 scans were pooled from this database.

5.2.2 ImUnity

ImUnity was used in this chapter as it was originally presented in [section 4.2](#). We trained three models, one for each axis and combined the three outputs using the median. Each model was trained on 3 subsets of the 1,698 images pooled from OASIS, SRPBS and ABIDE. OASIS and SRPBS data were randomly distributed between training and validation sets (respectively 80% (923 images) and 20% (230 images)). Then, in a 3-folds cross validation way, we added the ABIDE data subsets to each training, validation and test sets, i.e. 175 images each. We carefully controlled that training, validation and testing sets were perfectly independent ensuring that each harmonized image was produced by a model not trained or validated with the corresponding raw ABIDE image. For each cross-validation step, the three 2D models were trained in parallel for each direction, axial, coronal and sagittal respectively, using the same data subsets. Hyper-parameters were set as in [section 4.2](#), except for the biological loss which was set to zero as no biological feature was used during training (not available for both OASIS and SRPBS databases). In this study, we considered control subjects but also ASD patients present in the ABIDE database. This allows an observation of harmonization impacts on both populations.

5.2.3 ComBAT

The ComBAT method was used in the same way as in [subsection 3.2.2](#). That is to say, it was used independently for every voxel location, therefore harmonizing directly the image's voxels intensity. It is very frequent in the literature that ComBAT is used to harmonize derived metrics like radiomics ([Acquitter et al., 2022](#)), volumes or thickness ([Fortin et al., 2018](#)). However this practice has the defect of not harmonizing the images, thus requiring harmonization for every metric of interest. Furthermore, our use of ComBat is closer to ImUnity's process.

5.2.4 Data preprocessing

As for previous studies ([chapter 3](#), [chapter 4](#)), we used Robex [Iglesias et al. \(2011\)](#) for brain extraction and N4Bias [Tustison et al. \(2010\)](#) for intensity inhomogeneities correction. MR images were first co-registered to the publicly available and age specific 152-MNI templates [Sanchez et al. \(2012a\)](#). Then, White-Stripe normalization [Shinohara et al. \(2014\)](#) was used to align white matter peaks across all subjects (each WM peak was aligned to 0.7 after rescaling the whole image between [0:1]). After visual inspection with ROBEX to detect artifacts, we eventually included 1072, 81 and 524 T1w scans from OASIS, SRPBS and ABIDE databases respectively. All tissues and cortical or subcortical volumes were extracted using FreeSurfer [Fischl \(2012\)](#) from original and harmonized images (see [subsection 5.2.5](#)). The Desikan-Killiany atlas [Desikan et al. \(2006\)](#) was considered for cortical volume extraction.

5.2.5 Experiments

As first introduced in [section 5.1](#), we focus in this chapter on the impact of harmonization on brain development. We used 271 controls and 253 ASD ABIDE subjects between 5 and 25 y.o. to evaluate the impact of ImUnity on volume and thickness evolution with age in different ROIs. Age distribution of the selected ABIDE subjects is presented in [Figure 5.1](#). The multi-centric ABIDE database has already been shown to be impacted by site and scanner induced biases ([chapter 3](#), [chapter 4](#)), leading to poor results when pooling all data ([Sherkatghanad et al., 2019](#)). Assuming that data harmonization will help recovering expected developmental trajectories, we referred to several literature trends already presented in [chapter 2](#) ([Ducharme et al., 2016](#); [Lenroot and Giedd, 2006](#); [Vijayakumar et al., 2016](#); [Wierenga et al., 2014](#)) to evaluate ImUnity and ComBat effects. These trends were gathered from large mono-centric studies, preventing any harmonization related issue. The effect of harmonization was defined by its impact on volume evolution trends, and how harmonization moved them closer to those reported in literature. For this experiment, we considered separately healthy controls and subjects with ASD because ASD has significant impact on cortical and subcortical brain regions development during early childhood (see [subsection 2.1.2](#)).

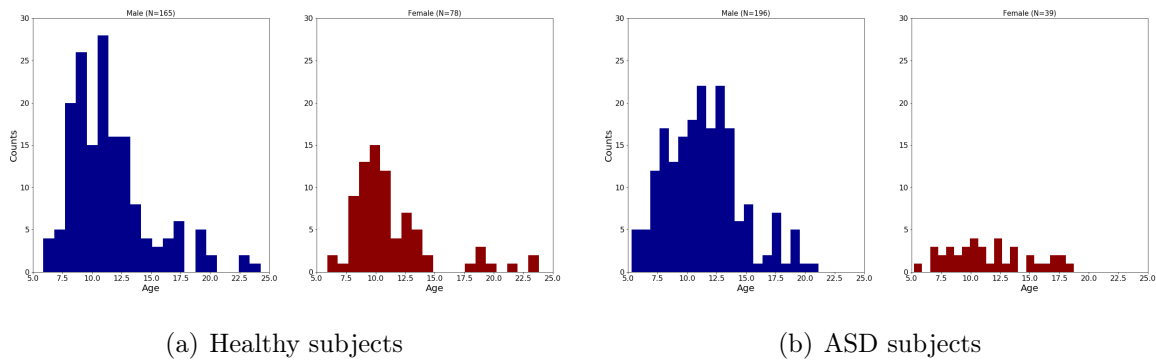


Figure 5.1 – ABIDE subjects age distribution between 5 and 25 years old for each sub-population considered during the experiments.

• VOLUME AND THICKNESS EVOLUTIONS: VISUAL COMPARISON TO THE LITERATURE

In a first investigation, we selected 80 brain ROIs, based on the Desikan-Killiany atlas, with known development evolution with age. Extracted metrics (here volume and thickness) were regressed out with age, leading to a linear or a quadratic evolution based on published literature results. Considering biological covariates effects like the sex of the status (ASD vs. healthy in our case), we considered independently each sub-population. We chose regions with large differences in size to better understand how harmonization corrected global (e.g. frontal or parietal lobes) or local (e.g. putamen or hippocampus) regions. For comparison purposes, ComBAT was also run on similar data. In a classical way, we provided to ComBAT images and site affiliation information as well as all biological features available (sex, age and status).

In this first experiment, we only considered visual evaluation of the trends evolution before and after harmonization, compared to the literature reference curves. Our hypothesis was that harmonization should bring observed trends closer to the literature

ones. First, we considered only volume trends for control subjects and then included ASD subjects to observe if the clinical status influences harmonization quality. In this experiment, we could not include ASD females as not enough literature results could be gathered. Similarly, for brain ROIs thickness evolution, only males (healthy and ASD) were considered. An example of the trend found for the total gray matter volume evolution in healthy males is given in Figure 5.2. Original data are indicated in red and seem to increase with age. This is clearly not in line with the reference data found in the literature and indicated in black. Both harmonization methods (ImUnity in green, Combat in blue) modify the tendency and seem to bring it closer to the reference.

- INTRODUCTION OF A QUANTITATIVE METRIC TO EVALUATE HARMONIZATION IMPACT.

Although the visual results provided in Figure 5.2 seem to indicate that both harmonization methods have a positive impact on the trends, it is difficult to conclude if one is better than the other. Initially (red line), the tendency is inverted compared to the literature. It is corrected after ComBat (blue line) but is ‘less inverted’ after ImUnity (green line). However, it is also possible that Combat corrections were too strong.

In order to be able to compare the methods more accurately, we proposed to compute a derived metric from the equation of a given observed volume trend and its equivalent reference equation found in the literature. The idea was to obtain a value reflecting the trends similarity between both equations. This metric is defined as follows:

$$H_{metric} = \int_{5y}^{25y} |f'(x) - g'(x)| dx \quad (5.1)$$

Where f corresponds to the reference equation of the ROIs and g is the one observed.

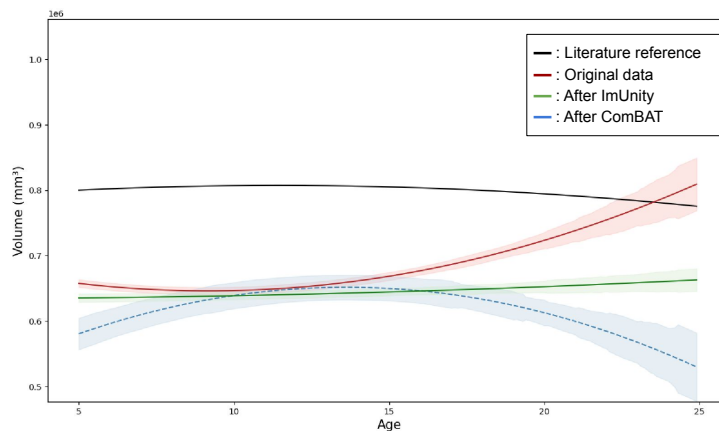


Figure 5.2 – Total Gray Matter volume evolution. In black the literature trends; respectively in red, blue, green the trends before harmonization, after ComBAT and after ImUnity.

The formula returns a value reflecting the ‘distance’ of the observed trend to the literature one. A value of zero corresponds to a perfect match of the two trends, while a high value will mean that the trends are different. Using this metric allows us to consider only the trends and not the shifts between the equations (the derivative is used). In fact, a shift between the literature and the observed equation is expected, mostly due to differences in hardware between studies (supposed to have a constant impact for every subject independently). However, there should not be trend differences, as the observed

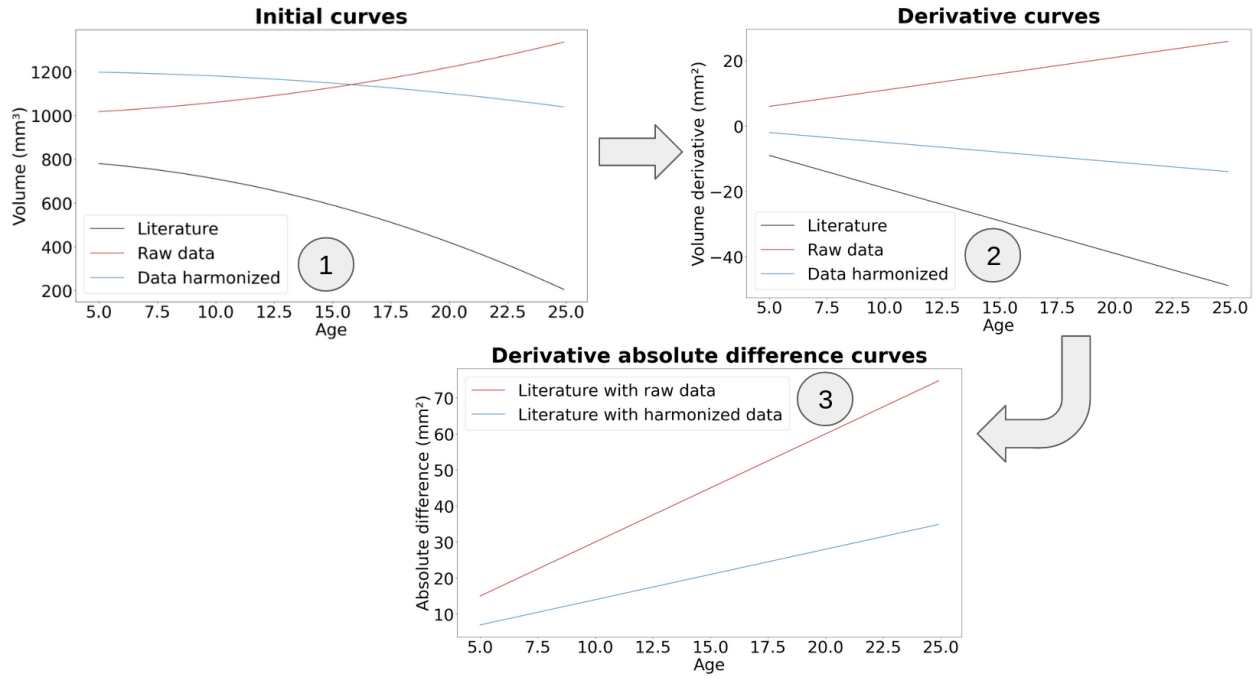


Figure 5.3 – Harmonization metric workflow. Starting from the equation, and following Equation 5.1 one can compute the metric value for each type of data (before and after harmonization). 1) Observed trends; 2) Derivative results for the observed trends; 3) Absolute difference results between the curve of interest (raw data or after harmonization) and the literature.

population are similar (healthy subjects from the same age range). Figure 5.3 illustrates the workflow to obtain the presented score, used to compute the metric for each data type and each ROI. Our hypothesis is that the metric value should get closer to 0 after harmonization as we expect harmonization to bring trends closer to the literature by removing unwanted site effect variations.

Due to some important volume variance between ROIs (which directly influences the order of magnitude of the metric), we eventually computed the ratio $\frac{Metric_{after\ harmonization}}{Metric_{original\ data}}$ to compare the methods. Doing so, we expected to have a ratio lesser meaning that harmonization had a positive effect and moved the trend closer to the literature. In the previous example Figure 5.2, we have a metric value of 202,986 before harmonization, 59,114 after ImUnity and 114,852 after ComBAT. This gives us a ratio of 0.29 for ImUnity, and 0.57 after ComBAT. As expected, we have a positive impact of both methods on this ROI volume (both ratios < 1). However, we have a smaller ratio for ImUnity, mainly due to the fact that ComBAT inverted too much the tendency. This observation is coherent with what we can observe when giving a close look at Figure 5.2. This metric can be used to quantitatively compare harmonization effects, independently for every ROI. Similarly to the first experiment based on visual inspection, we can observe positive or negative harmonization effects based on the metric value. In addition, it can also be used to quantify the need of harmonization for the ROI considered. This is an important point as some regions might need more correction than others. To do so, for each ROI, the associated metric is divided by its mean volume between 5 and 25 y.o. This is motivated by the fact that it is not possible to directly compare the

metrics as their orders of magnitude vary significantly across all the ROIs. Dividing the original metric by the ROIs' mean volume leads to a metric without unit that can be used to quantify the need for harmonization. This metric can be used to spatially visualize harmonization needs. As one might expect non-homogeneous induced site noises, resulting in some brain areas more altered than others. Additionally, this information can also be used to observe any changes of harmonization performance regarding its original need. We expect to have a more significant positive effect (ratio < 1) for ROIs mostly concerned by harmonization. On the other hand, ROIs trends already in line with the literature are less likely to have such a ratio. Ratios should also be lower than 1 as harmonization should not degrade the trends.

- **MIXED LINEAR MODEL ANALYSIS**

In another type of experiment, inspired by [Beer et al. \(2020\)](#), a linear mixed effect model analysis was run on volume and thickness estimates, considering all available features: age, sex, status, site. Then, by considering site affiliation as a cluster factor, we could estimate its impact on the estimates (volumes or thickness). This was done with the Kenward-Roger (KR) tests ([Kenward and Roger, 1997](#)), testing for estimates and images origin joint significance and using the pbkrtest R package ([Halekoh and Højsgaard, 2014](#)). In other words, a statistical-test was done for each parameter under the hypothesis H_0 : "the parameter has no influence on the observed estimate (volume or thickness)". For each test, according to its associated p-value, we would reject ($p < 0.05$) or not H_0 ($p \geq 0.05$). Our hypothesis was that the groups (site affiliation) would have a significant impact ($p < 0.05$) on brain development before harmonization and this effect would vanish after harmonization ($p \geq 0.05$). Site effects are likely to be constant and to affect every patient in the same way. In fact, these noises affect the images independently of the patient and his/her biological features. On the other hand, we expect to have an increase of biological features significantly correlated to the estimates. Note that this part is difficult to analyze, as every ROI in the brain is not affected in the same way by these biological features and harmonization must not force biological correlations.

- **SURFACE-BASED GROUP ANALYSIS - FREESURFER**

Finally, the last experiment consisted in a **surface-based thickness group analysis** on ABIDE healthy controls, using site origin as groups and integrating biological covariates (age and sex) in our model. Surface thickness and group analysis were obtained with Freesurfer. Cluster-wise correction ([Hagler et al., 2006](#)) was used with a threshold of 10^{-3} as recommended in [Greve and Fischl \(2018\)](#) to avoid false positive clusters. This procedure allows to visualize groups impacts (here site affiliation) on brain development, and to visualize brain areas most affected by these variations and to observe the effects on these clusters. This comes in complement to the previous experiment based on the derived metric used to quantify the need of harmonization in different ROIs. However, no brain atlas parcellation was used here and the brain surface was considered as a whole. The hypothesis here is that this procedure should detect some surface clusters significantly influenced by site affiliation for original data and none after harmonization. Also note that harmonization must not induce new clusters.

5.3 Results

5.3.1 Volumetric analysis on healthy subjects

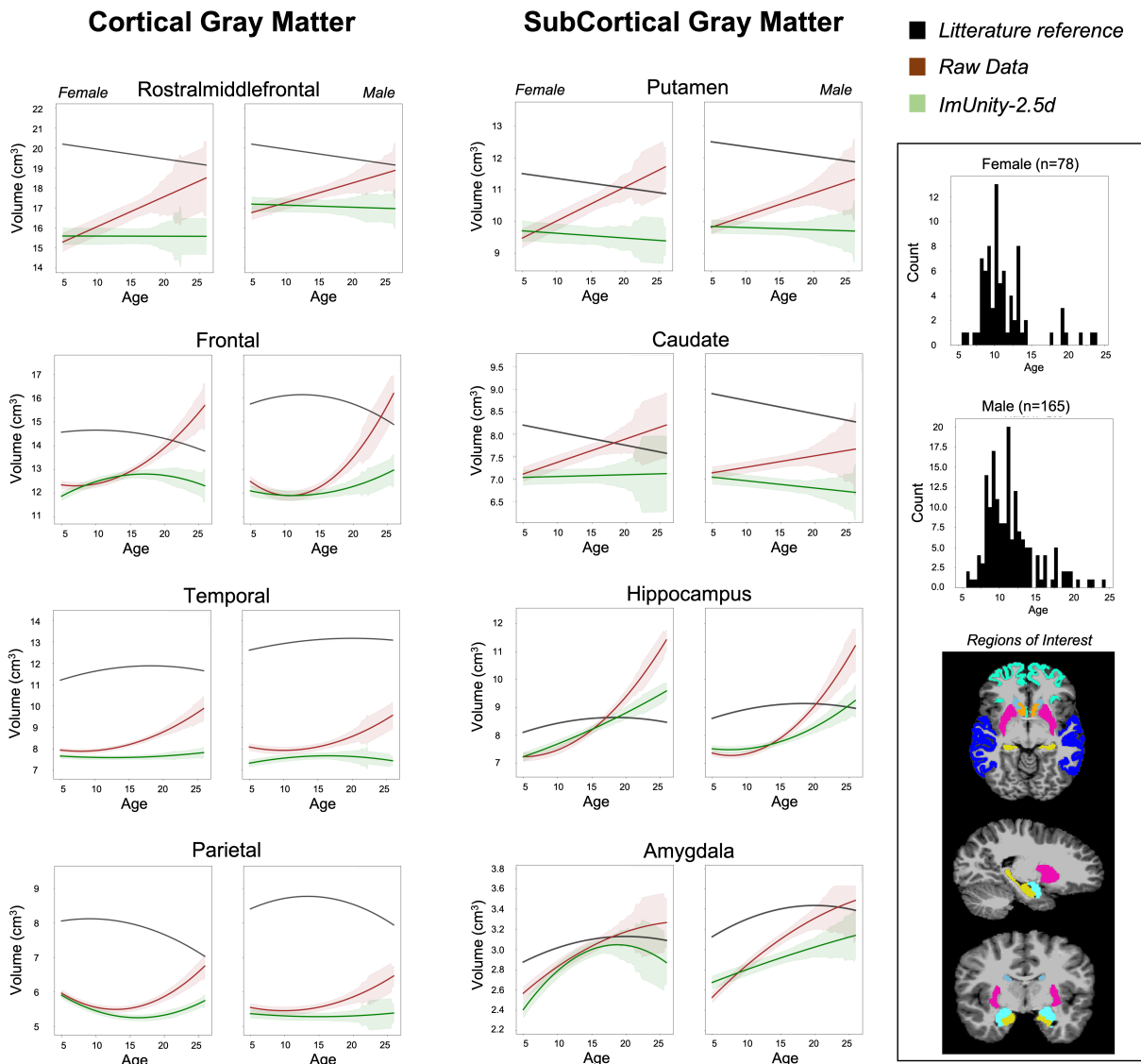


Figure 5.4 – Effect of ImUnity on the evolution with age of cortical (left) and subcortical (right) gray matter regions volume, for healthy males and females independently. Black lines correspond to literature trends while green (resp. red) lines refer to trends found on ImUnity-harmonized (resp. without harmonization) data. 95% confidence intervals are displayed for both without and with ImUnity harmonization.

Figure 5.4 presents the evolution with age (linear or quadratic interpolated trends with associated 95% confidence interval) of the volume of different cortical and subcortical structures when considering only healthy subjects. The age distributions are provided for males and females, centered around 10 years old with less data above 17 years old. In almost every region, before harmonization (red curve), there is a trend for the volume to increase with age which is clearly not coherent with the literature (black curve). Although the total brain

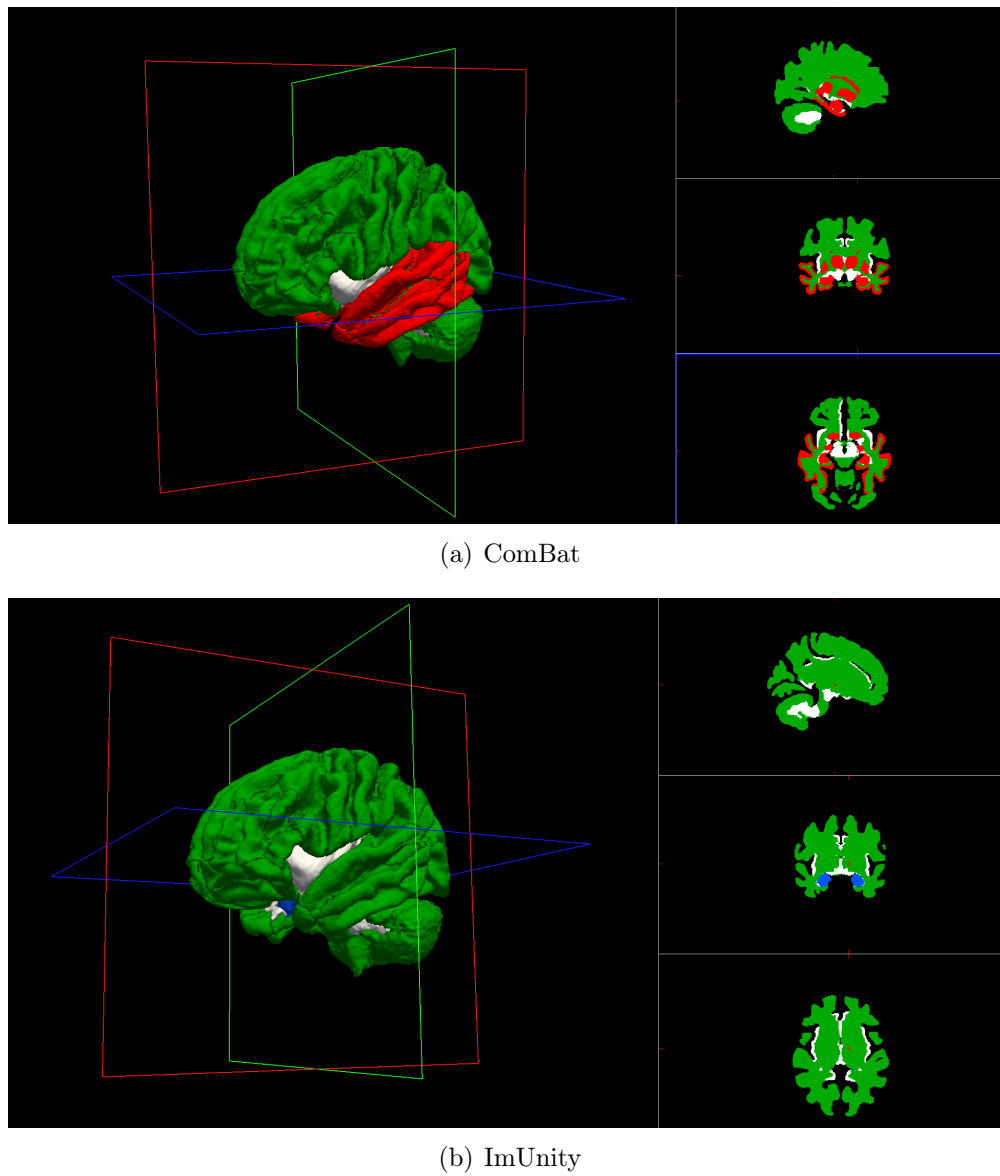


Figure 5.5 – Average harmonization methods impact on brain regions volume for healthy females. **A)** Effect of ComBAT, **B)** Effect of ImUnity. Red: negative effect of harmonization ($\text{ratio} > 1.1$); Blue: no noticeable effect ($0.9 > \text{ratio} > 1.1$); Green: positive effect of harmonization ($0.9 > \text{ratio}$); White: non-investigated.

volume is known to increase during this age range, it should not be the case for every ROI's volume. For example, as seen in [subsection 2.1.1](#), the total GM volume is known to decrease while the WM tends to increase. For every ROI presented in [Figure 5.4](#), harmonization with ImUnity (green curve) improves trends results and brings them closer to the reference values (see for example hippocampus or amygdala). In some regions ImUnity reverses the original trend (e.g. putamen), changing from linear increase to linear decrease. In frontal (female) or temporal (male) lobes, ImUnity changes the trend from quadratic increase to quadratic decrease. The harmonization results do not depend on the size of the region as similar findings can be found in small subcortical (e.g. caudate) or large cortical (e.g. frontal and temporal lobes) regions. We note that harmonization does not perfectly correct all structure trends.

In certain regions (e.g. parietal lobe), the trend is improved but not perfectly aligned with the literature.

Table 5.1 presents the results of ImUnity and ComBAT on the metric introduced in subsection 5.2.5. It highlights the positive impact of both methods on volumetric trends for healthy male and female subjects and confirms the previous results. To be more precise, we can observe that ImUnity does not degrade trends whereas ComBAT seems to be less adapted for some specific ROIs, especially for female data-harmonization. We can also note the robustness of ImUnity for gender, as it presents very good results for both, whereas ComBAT has difficulties with female data. This is highlighted by Figure 5.5, 5.6(a) and 5.6(c), showing the ROIs (temporal lobe, accumbens) for which ComBAT struggles while ImUnity seems to have had a global positive impact.

Based on previous results, and highlighted by Fig. 5.6(e), we can observe that the ‘lateral ventricles’ and the ‘corpus callosum’ volumes are the most concerned by harmonization. To a lesser extent, the volumes of the ‘brain stem’, the ‘white matter’, the ‘thalamus’ and the ‘pallidum’ also diverged greatly from the literature. The evolution of the trends of all these volumes were corrected through harmonization. Finally, we can point out the minimal effect of combat on ‘accumbens’ and ‘temporal lobe’ volumes while ImUnity seems to have improved volume trends for all regions with a less noticeable effect for the ‘amygdala’.

	Male		Female	
	ComBAT	ImUnity	ComBAT	ImUnity
Metric ratio < 0.9	85%	95%	53%	94%
Metric ratio > 1.1	15%	0%	41%	0%

Table 5.1 – Harmonization volume metric ratio for ComBAT and ImUnity, on healthy subjects. As presented in Section 5.2.5, a ratio < 1 corresponds to a volume trend alignment after harmonization towards the reference trend. Here we present the percentage of ROIs with ratio < 0.9 (improvement) or > 1.1 (degradation), ratios in between are considered as ‘untouched’.

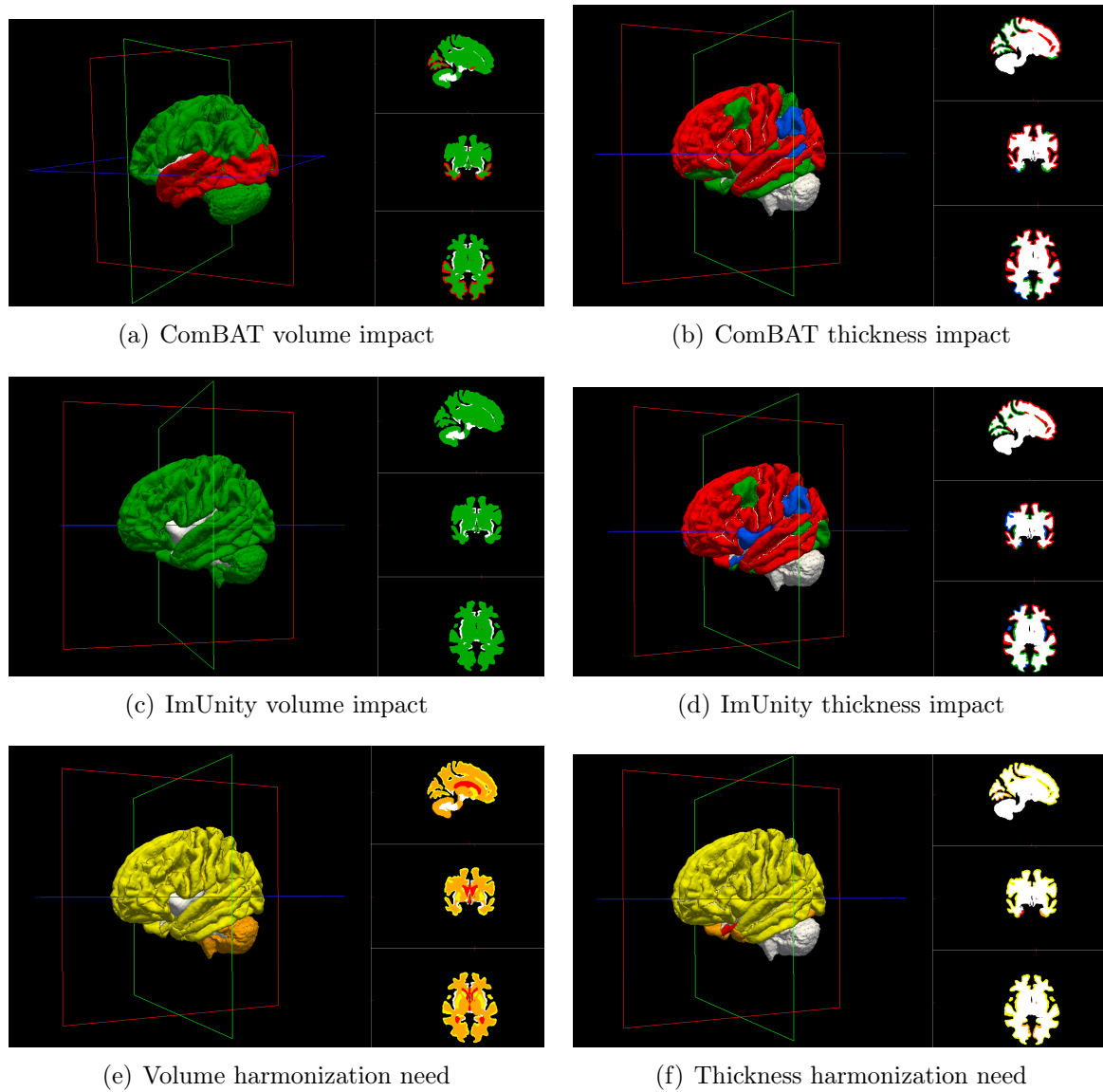


Figure 5.6 – Thickness and volume harmonization impacts for healthy males from the ABIDE database.

A) Effect of ComBAT on volumes metric; **B)** Effect of ComBAT on thickness metric; **C)** Effect of ImUnity on volumes metric; **D)** Effect of ImUnity on thickness metric;

Color code for A-D) Red: negative effect of harmonization ($\text{ratio} > 1.1$); Blue: no noticeable effect ($0.9 > \text{ratio} > 1.1$); Green: positive effect of harmonization ($0.9 < \text{ratio}$); White: Non investigated;

E) ROIs volume metric need for harmonization; **F)** ROIs thickness metric need for harmonization.

Color code for E-F) Yellow: ROIs needing less harmonization ($0 < \text{ratio} < 0.5$); Orange: ROIs requiring harmonization ($0.5 < \text{ratio} < 1$); Red: ROIs requiring strong harmonization ($1 < \text{ratio}$); White: non investigated.

5.3.2 Thickness analysis on healthy subjects

Figure 5.7 presents harmonization impacts on thickness development across the age range 5 to 25 y.o. In contrast to volume results (Figure 5.4), it seems that harmonization did not have such a positive impact. For some ROIs (Cuneus, Medial orbito frontal), ImUnity even degraded thickness trends. Table 5.2 confirms this point. Although harmonization was beneficial for about one-third of the ROIs considered, it shows that for nearly 50% of them, both harmonization solutions worsened the observed trends. This result goes against the initial harmonization hypothesis and requires more investigations.

Fig. 5.6(b) and Fig. 5.6(d) visually presents the effect of both solutions on ROIs investigated. Once again we can observe that harmonization did not go as well as expected. This can first be explained using Fig. 5.6(f), where we can clearly observe a lesser need for harmonization for the thickness than for the volumes. This is also illustrated in Fig. 5.6(f), in which we can spot the difference between volume and thickness evolution needed for harmonization. In fact, harmonization seems to be more needed when considering ROIs volumes than thickness. This point could explain previous results on thickness harmonization as some ROIs did not require harmonization and thus it was more likely to worsen further extracted metrics.

Based on these figures, it seems that the ‘temporal’ and ‘occipital’ lobes are the main two regions concerned by thickness harmonization. Even in these two regions, ComBat could not correct the metric trends whereas ImUnity could improve ‘occipital lobe’ thickness trend only.

	Male	
	ComBAT	ImUnity
Metric ratio < 0.9	36%	31%
Metric ratio > 1.1	46%	49%

Table 5.2 – Harmonization volume metric ratio for ComBAT and ImUnity, on healthy males. As presented in Section 5.2.5, a ratio < 1 corresponds to a volume trend alignment after harmonization towards the reference trend. Here we present the percentage of ROIs with ratio < 0.9 (improvement) or > 1.1 (degradation), ratios in between are considered as ‘untouched’.

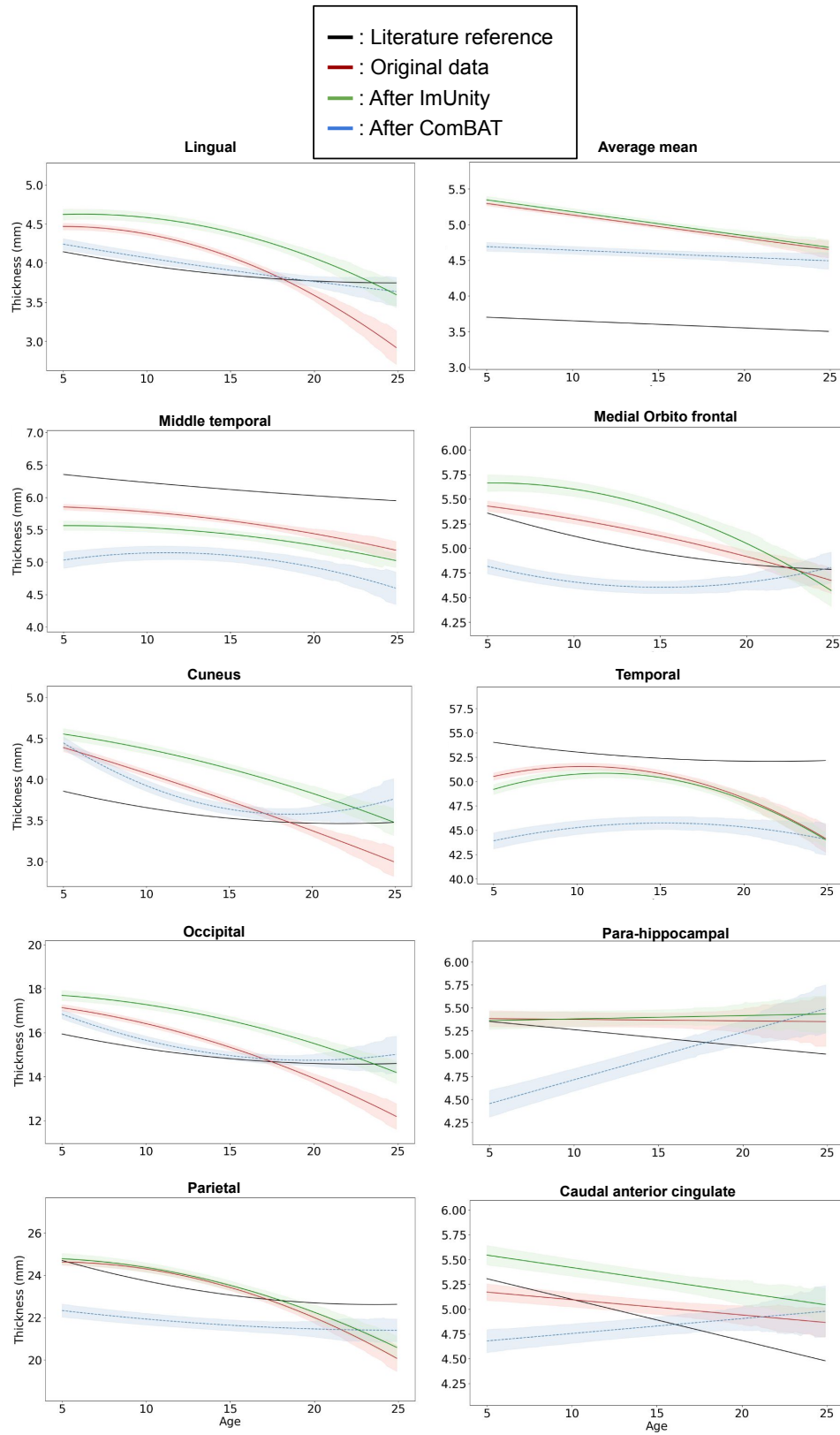


Figure 5.7 – Effect of ImUnity and ComBAT on the evolution with age in considered brain ROIs’ thickness, for healthy males. Black lines correspond to literature trends while green (resp. blue) lines refer to trends found on ImUnity-harmonized (resp. ComBAT-harmonized) data. Red lines correspond to trends obtained on original data before harmonization. 95% confidence intervals are displayed.

5.3.3 Impact on ASD patients

So far, we have only presented results for healthy subjects. In this section, we present the impact of harmonization in the case of ASD patients. As mentioned in [subsection 2.1.2](#), we only present results for ASD males as we could not gather enough literature results for ASD females brain development.

[Table 5.3](#) presents our metric ratio for ROIs volume and thickness. Similarly to previous results on healthy subjects, we observed a clear positive impact of harmonization on ROIs volumes for about 90% of ROIs both methods could bring their trends closer to the literature. On the other hand, it was less beneficial for ROIs thickness, and in nearly 40% of investigated ROIs, the harmonization was negatively impacting the trends. [Figures 5.8\(e\) and 5.8\(f\)](#) confirm these results and highlight harmonization benefits for ROIs volume metric whereas its effect is less clear for the thickness. Similarly to [subsection 5.3.2](#), this can be explained by a lesser need for harmonization for the thickness as illustrated in [Figure 5.8](#).

	Volume		Thickness	
	ComBAT	ImUnity	ComBAT	ImUnity
Metric ratio < 0.9	94%	88%	53%	53%
Metric ratio > 1.1	6%	6%	45%	37%

Table 5.3 – Harmonization volume & thickness metric ratio for ComBAT and ImUnity, on ASD males. As presented in [Section 5.2.5](#), a ratio < 1 corresponds to a volume trend alignment after harmonization towards the reference trend. Here we present the percentage of ROIs with ratio < 0.9 (improvement) or > 1.1 (degradation), ratios in between are considered as ‘untouched’.

[Figure 5.8](#) can also be used for harmonization needs comparison between healthy and ASD subjects. For example, we found a similar harmonization need for the volumes (lateral ventricles, white matter and thalamus). However, for the thickness, results seem to be impacted by the subject’s status. While the thickness of temporal and occipital lobes were concerned by harmonization for healthy males, it was not the case for ASD subjects for which the frontal lobe thickness were the metrics the most affected by sites induced noises.

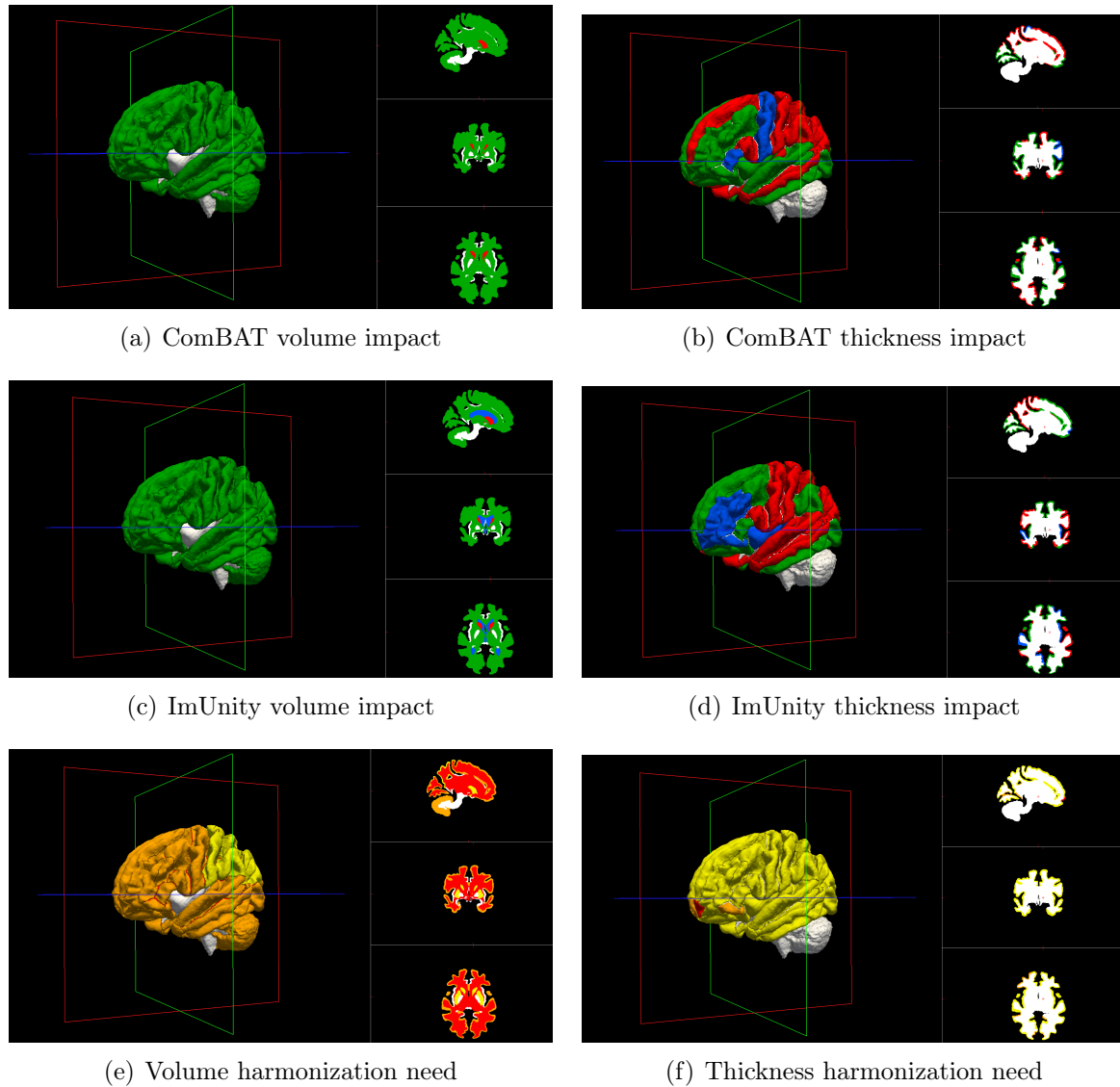


Figure 5.8 – Thickness and volume harmonization impacts for ASD males from the ABIDE database.

A) Effect of ComBAT on volumes metric; **B)** Effect of ComBAT on thickness metric; **C)** Effect of ImUnity on volumes metric; **D)** Effect of ImUnity on thickness metric; Red: negative effect of harmonization ($\text{ratio} > 1.1$); Blue: no noticeable effect ($0.9 > \text{ratio} > 1.1$); Green: positive effect of harmonization ($0.9 < \text{ratio}$); **E)** ROIs volume metric need for harmonization; **F)** ROIs thickness metric need for harmonization.

Yellow: ROIs needing weak harmonization ($0 < \text{ratio} < 0.5$); Orange: ROIs requiring harmonization ($0.5 < \text{ratio} < 1$); Red: ROIs requiring strong harmonization ($1 < \text{ratio}$); White: non-investigated.

5.3.4 Mixed Linear Model analysis

As mentioned in [subsection 5.2.5](#), we eventually ran statistical tests to observe the impact of harmonization on biological and acquisition features. [Table 5.4](#) and [Table 5.5](#) present the

number of tests for which H_0 was rejected ($p < 5.10^{-5}$), independently for volume and thickness metrics investigation. Except for the last line of the table, it was expected to reject H_0 as these biological features directly impact brain development. This is different for the last line of the table which focuses on the site affiliation correlation. For this case, the idea was not to reject H_0 as site affiliation should not impact brain development. The reported number is expected to be low.

Table 5.4 highlights the need for harmonization as the number of correlated volumes to sites affiliation is fairly high (81%). As expected, this number drops after harmonization (15 after ImUnity and 9 after comBAT), signifying the positive effect of harmonization. On the other hand, results for biological features are less encouraging, as in most cases, the number of correlated volumes decreases after harmonization. This is mainly the case for ComBAT which seems to have removed the variations induced by the status (here being ASD). ImUnity was able to preserve most of the biological variation and was able to increase the number of volumes correlated to age (26 \rightarrow 29).

Table 5.5 presents the equivalent results for ROIs thickness metrics. Here we can first observe that brain thickness development is less impacted by harmonization issues, as only 2% of brain regions (frontal and cingulate areas) have their thickness significantly impacted by site affiliation. This is in line with the previous results (see subsection 5.3.2). After ComBAT, only one remaining region (the frontal lobe) seems to be impacted by site effects. On the other hand, ImUnity seems to have had a negative effect as 5 regions thickness are significantly impacted by site effects after harmonization (the cingulate, temporal, and frontal areas). In addition, both harmonizations could not improve biological correlations. Note that except for ImUnity and the age, both methods seem to have induced a global loss of biological information in the data as the number of correlated thickness evolution reduced after harmonization.

	Original data	ImUnity	ComBAT
Age	26	29	17
Sex	8	4	3
Status	29	16	0
Age ²	28	24	9
Age.Sex	12	4	2
Age.Status	34	18	0
Sex.Age ²	19	5	1
Status.Age ²	35	23	0
Site affiliation	73	15	9

Table 5.4 – Number of KR tests rejecting H_0 : "the feature has no influence on the observed volume" over the 80 ROIs investigated.

	Original data	ImUnity	ComBAT
Age	84	55	17
Sex	17	6	9
Status	12	4	5
Age ²	38	16	17
Age.Sex	7	10	10
Age.Status	19	5	6
Sex.Age ²	5	6	9
Status.Age ²	56	6	9
Site affiliation	3	5	1

Table 5.5 – Number of KR tests rejecting H0: "the feature has no influence on the observed thickness" over the 124 ROIs investigated (independently for left and right hemispheres).

5.3.5 Freesurfer group analysis

Our last experiment consisted in a surface based group analysis ran through the Freesurfer software. For this experiment, we only considered healthy subjects. We also considered sex as a biological covariate of interest in order to avoid cluster effect due to sex related differences.

Figure 5.9 presents the obtained results, showing in yellow the clusters detected by the model that are significantly impacted by site affiliation. As for previous experiments, results before and after harmonization are illustrated. It is interesting to point out the positive effect of harmonization. Three clusters were identified on original data, only one after ComBAT and none could be detected after ImUnity. Once again, this highlights the need for harmonization and also confirms the site removal effect of both methods. Note however that, similarly to the experiment run in section 4.2.4, this result is necessary but not sufficient. In fact if we refer to our definition of harmonization ('removing site or scanner related noises while preserving biological information in scans'), this experiment only verifies the first point.

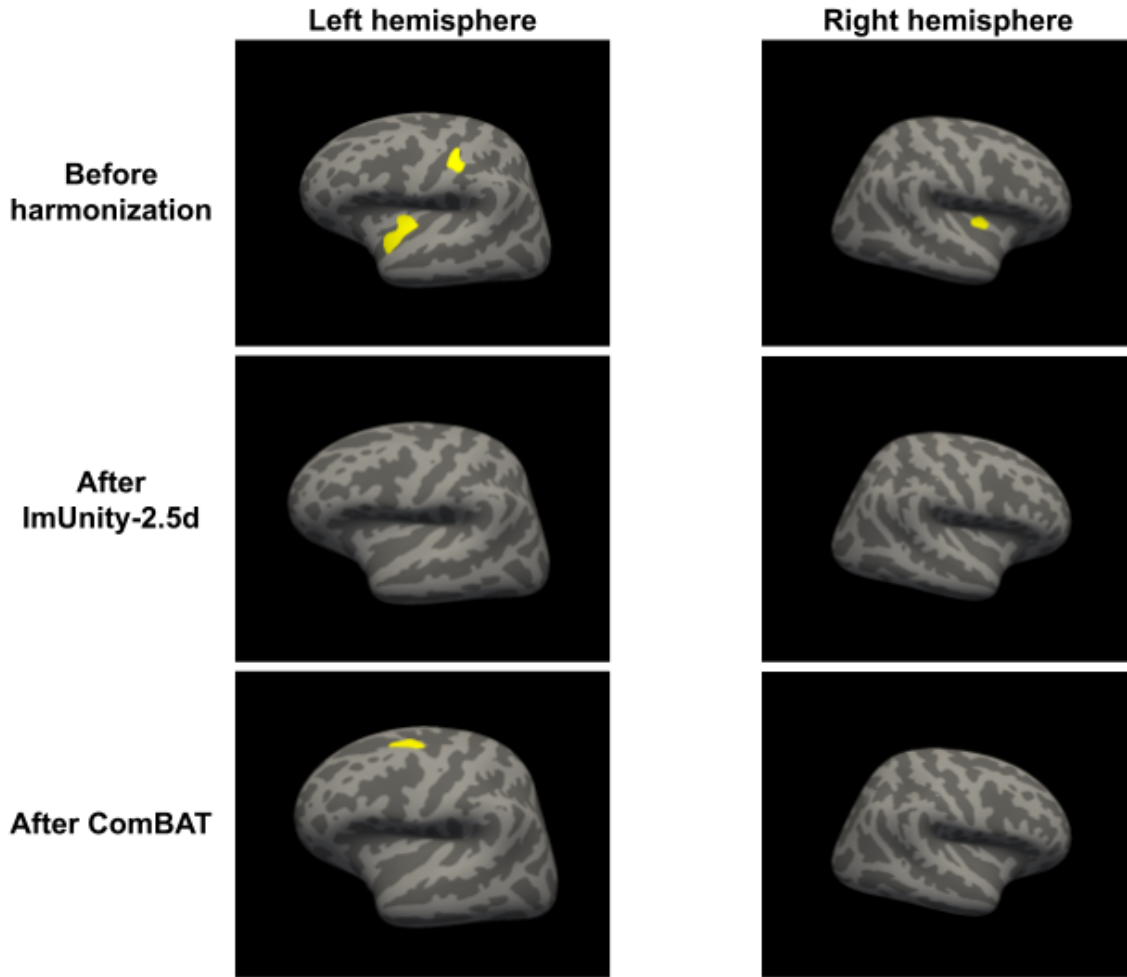


Figure 5.9 – Surface-based thickness group analysis results for both hemispheres, on raw data (top row), after ImUnity harmonization (middle row) and after ComBAT (bottom row). Spots in yellow indicate clusters areas impacted by site or scanner affiliation, accounting for biological (sex and age) covariates.

5.4 Discussion

In this chapter, we further evaluated our proposed harmonization solution. The first validation steps using traveling subjects, site and patients classifications ([section 4.3](#)) already demonstrated the potential of our method. This chapter completes these previous works and demonstrates the positive effect of ImUnity on brain development trajectories. Experiments were also ran on ComBat-harmonized data for comparison purposes. Using a derived metric ([Equation 5.1](#)) to quantify trend variations, we were able to assess the effect of both solutions. It was also helpful to visualize the need for harmonization of each ROI taken independently.

Regarding brain volumetry, ImUnity demonstrated its ability to remove site or scanner biases, while having a positive effect on the study of cortical and subcortical developmental trajectories. ImUnity and ComBAT could adapt images contrasts while preserving geometrical information. In our first experiment, we showed that this is sufficient to improve the apparent trends in brain volume evolution and bring them closer to the results from large mono-centric studies. This was confirmed by the metric ([Equation 5.1](#)), with a global

improvement for volume tendency. Most of the ROIs volumes seemed to be impacted by harmonization issues, and we could improve about 95% (resp. 88%) ROI's volume evolution among healthy subjects (resp. ASD males). Note that all these results are based on the derived metric proposed to compare ROI development tendencies to the literature. No statistical test was run for this part, but these results give an interesting data interpretation and a good intuition for harmonization impacts. Moreover, these results were confirmed by the linear mixed model approach, a well known statistical model that can be used to test for significant features effect. In our case, it was used to model the impact of all available features (age, sex, status and origin) plus some combinations of them up to the order 3 (age^2 , $\text{sex} \cdot \text{age}^2$, $\text{status} \cdot \text{age}^2$) on ROIs volume evolution. It allowed to highlight the need for harmonization as the origin (here the site) of images was significantly impacting most of ROIs volume trend before harmonization. This was reduced after harmonization. On the other hand, harmonization seemed to have reduced biological correlations, which goes against our initial hypothesis. Both methods showed similar results but our method seemed more adapted as ComBat removed more biological significant correlations and corrected less ROI volume trends than ImUnity. It is however important to note that it is not clear from the literature review that every ROI should be significantly correlated to every biological feature taken into account in this study. For example, several studies ([Vijayakumar et al., 2016](#); [Wierenga et al., 2014](#)) present either linear or quadratic correlation with age depending on the brain region considered. Some reports also show contradictory results in specific regions of interest. In order to be able to conclude, a thorough analysis of the trends found in various mono-centric studies still needs to be performed and the results compared to the ones found in our statistical analysis.

We also analyzed the impact of harmonization on brain thickness development. Results were less encouraging as harmonization was less beneficial on the trends evolution with age. We could even point out several cases for which harmonization degraded the tendencies. Our results are not in line with previous studies on the topic ([Beer et al., 2020](#); [Fortin et al., 2018](#)). However, it is important to note the different way of using ComBat in our study where we harmonize images intensity first and then extract the metrics of interest (here brain volumes and thickness) while the metrics are used directly in ComBat in the previous studies. We used this pipeline in a realistic clinical situation, in which harmonized images could contribute to the analysis of different clinical assumptions. But these different usages of ComBat might explain the differences between the results. Also note that further investigations using our derived metric ([Equation 5.1](#)), and the linear mixed model approach also seemed to point to large differences in the need for harmonization between brain volumes and surface time evolution.

Another interesting point to mention is the crucial role of image contrast in standard image analysis workflows. Indeed, our studies show that both harmonization methods only modify images contrast and preserve the anatomies (see [section 4.3](#)). Yet, these induced variations directly impacted Freesurfer (which is a widely used software) extraction metrics. Thus, as the domain of acquisition highly impacts image contrast, it is likely to also impact further analyses using Freesurfer or any other such software.

In our work, the reference contrast for the harmonization process was chosen arbitrarily. Yet, ImUnity could be used to harmonize all images to any of the reference domains. It could thus be interesting to see how this choice impacts the results. It is very likely that there is an optimal contrast domain for a specific application, as a radiologist would adapt the observed contrast to best see a specific image feature. As the chosen reference contrast has an

impact on Freesurfer metrics extraction, it is very likely that this could explain the difficulties encountered regarding the thickness trends. In addition, it could be interesting to enlarge the age range by adding sites from other databases (e.g. UK Biobank, [Sudlow et al. \(2015\)](#)) and see how ImUnity generalizes the results knowing that it has already been trained on images from adult volunteers of the OASIS and SRPBS databases. Given these encouraging results, ImUnity could be used to look at the brain volume trajectories in patients with brain diseases. In this case, it could also be easily adapted to other types of MRI sequences such as diffusion and resting-state fMRI acquisitions.

Chapter 6

Conclusion and perspectives

This study was initiated by the French national project DEFIDIAG, focusing on Intellectual Disorders (ID) diagnosis among children, and was supported by MIAI, the Grenoble Multidisciplinary Institute in Artificial Intelligence, within the multiomics chair. In a context involving a large multi-centric database gathering data from children with a large age range, several challenges had to be addressed.

The first concern about the DEFIDIAG project was the population age range and its consequences on further analyses. After a bibliographic review about brain development, we assessed the importance of biological covariates such as age, sex or ID on brain trajectories. Moreover, we were able to set up a coherent pre-processing workflow in order to homogenize the data while preserving the biological variations induced by age. This was mainly done by using an open access age-specific templates database, representing average brains at 6 months intervals.

The second concern was MR data-harmonization between the multiple sites of acquisition and has become our main point of interest. Harmonization is a very active research field and has gained visibility due to the increase in the number of multi-center databases such as DEFIDIAG, ABIDE, ENIGMA, UKBIOBANK, ADNI, etc. As mentioned in [section 1.2](#), [Bottani et al. \(2022\)](#) highlighted important intra-hospital variations in a context of a clinical care study. It would be very interesting to evaluate the capacity of existing harmonization solutions to be included in routine clinical care. In these important multi-site databases, we have observed in our work the importance of taking into consideration the site induced noises during the analysis and predictions of ML models. As described in [chapter 2](#), several data harmonization solutions have been proposed since the late 90s. In [chapter 3](#), we proposed to compare a recent Deep-Learning (DL) based solution, ‘cycleGAN’, to ComBAT, a statistical solution considered as a reference in the domain. This study on real and synthetic data led to interesting results that emphasized the great potential of DL based approaches. On the other hand, even if cycleGAN provided good results, its architecture was limited to ‘two site harmonization’ only and was thus hard to use in most multi-centric studies scenarios. Other practical limitations were identified and we concluded that no existing harmonization solution was suited for our needs. Therefore, we proposed an original DL harmonization solution called ImUnity. In [chapter 4](#), we described this tool and presented complementary experiments to validate our approach. Harmonization on traveling subjects led to state of the art harmonization results. Using two classification tasks (1: site; 2: status), we showed that

ImUnity was able to remove site effects and preserve biological information. Additionally, it was shown that the solution could generalize its training to data coming from unseen sites. In a final study presented in [chapter 5](#), we tried to estimate the need and impact of MR harmonization in brain developmental analyses. This was done through multiple approaches that included: 1) an investigation on brain development evolution with age compared to literature results; 2) the introduction of a metric to compare trends before and after harmonization with reference trends from literature; 3) statistical tests consisting in a linear mixed model approach to better estimate harmonization impact on features (biological or not); 4) a group analysis effect ran to highlight brain surface areas mostly concerned by harmonization. Overall, we observed significant improvement of ROIs volume evolutions after harmonization with similar performance for healthy volunteers and ASD patients. The impact on thickness evolution was less significant and negative impacts of harmonization were even observed in some ROIs.

Several limitations of our work can be noted. First, our proposed solution is a 2.5D setup. Although it requires less parameters to fit than a 3D solution, it also tends to produce outputs of lesser quality. Second, ImUnity was only tested for harmonizing anatomical T1w images and we have not tested the model on other types of sequences. Theoretically, our method should produce similar results (as it was not designed specifically for T1w data), but it could be very interesting to evaluate its performances on MR diffusion images or images produced by CT scanners. Ultimately, ImUnity should also allow for multi-sequence harmonization (e.g. T1w and T2w). This is a real challenge that has not been investigated in the literature yet but could have a major impact on clinical practices. We have started a study on a combined T1w and CT harmonization, and the promising preliminary results warrant further investigations.

Additionally, geometrical distortions should be included as well as contrast variations. The last presented results ([section 5.3](#)) suggest that all ROIs are not impacted similarly and that contrast alignment was not enough to remove the entire site's induced signal. This goes against the global hypothesis that these variations only impact images contrast and not their geometry. To tackle this point, in the same way as in the first study ([chapter 3](#)), we could first introduce synthetic local geometrical distortions (using elastic transformations for example) and observe harmonization impact on it. In a second time, these variations could be introduced in ImUnity's training (alongside contrast variation) in order for the model to learn how to detect these local geometrical variations and to correct them.

Another crucial point to mention is the choice of the reference contrast used during ImUnity inference. Our last study highlighted the sensitivity of conventional software, such as Freesurfer, to contrast variations. One assumption was that there should be an optimal reference contrast for each task. A study to find the best reference contrast should be done. This would lead to better harmonization results, and probably significant improvements in brain cortical thickness estimates.

On top of technical developments, deeper investigations into brain developmental trajectories have to be made. While results were clear for brain ROIs volume evolution, the results for cortical thickness were insufficient to conclude. We note that the age distribution ([Figure 5.1](#)) of the subjects was non uniform, with a distribution close to a Gaussian curve centered around 15 years old. Adding subjects from other databases could homogenize the distribution for each population and reduce the confidence interval of the regressions. Simi-

larly, experiments could be run on different multi-centric databases presenting adult subjects with less biological variations. Another critical point, already mentioned in [section 5.4](#), is the investigation of biological impact on each independent ROI. Based on literature results from large mono-centric results, we should be able to better estimate the impact of each biological feature on the observed volumes or thicknesses and actually count the number of ROIs positively or negatively impacted by the harmonization methods.

A last point that would be helpful to better evaluate harmonization need and impact is the creation of a large ‘traveling database’ presenting important biological variations. This could focus on a specific pathology among children that would successively be scanned in several sites. If the number of subjects is large enough, we should be able to better estimate sites induced noises, biological information etc... This would allow an easier comparison of harmonization solutions but also a better understanding of the effects of each scanner on final images.

In conclusion, it seems that the tools developed during our project can be used to analyze the DEFIDIAG data when available. Even if new features can be added and further validations can be performed, ImUnity seems to be able to provide high quality images in all directions, remove site/scanner bias, and improve patients classification and brain developmental volume trajectories in both large and small brain regions. Our work has been submitted for publication in a peer reviewed journal and presented in several national and international conferences. As a consequence, collaborations with French institutes (CRE-ATIS in Lyon and ARAMIS in Paris) and US laboratories have been established to share and further test the potentials of our approach.

Chapitre 7

Résumé du manuscrit

Résumé

Ce projet de thèse s’inscrit dans le cadre du projet national français DEFIDIAG. Il s’agit d’un projet multi-centrique axé sur le diagnostic de déficience intellectuelle chez l’enfant. De ce contexte spécifique, plusieurs préoccupations ont émergé et ont été abordées au cours de cette thèse.

La première étude a porté sur le développement du cerveau, de la petite enfance au début de l’âge adulte. Ce manuscrit présente le développement normal du cerveau avec l’âge ainsi que l’impact des covariables biologiques comme le sexe ou la pathologie. Ces résultats sont cruciaux pour mieux comprendre l’impact des variables biologiques sur le développement du cerveau et pour pouvoir observer les changements induits par des pathologies d’intérêt. En parallèle, la prise en compte de ces variations lors de l’étape de pré-processing est essentielle. C’est pourquoi nous présentons les recommandations de la littérature concernant ces étapes.

En raison de la nature qualitative des acquisitions IRM, produisant des images pondérées c’est-à-dire non quantitatives, ce projet s’est rapidement concentré sur l’harmonisation des données IRM. Les séquences IRM, telles que T1w ou T2w, sont sensibles à la variabilité des scanners entre les fournisseurs (matériel, logiciel), aux choix techniques (par exemple, les paramètres de la séquence) et aux artefacts d’acquisition. Par conséquent, la mise en commun d’images provenant d’études multi-centriques afin d’aborder une question clinique ou biologique spécifique ne garantit pas une augmentation de la puissance statistique, en raison d’une augmentation parallèle de la variance non biologique. Pour ces raisons, une revue de la littérature portant sur les méthodes d’harmonisation a été réalisée afin de rassembler les solutions existantes et de comprendre les principaux défis restants.

Sur la base de notre analyse bibliographique, il nous est apparu que les solutions d’harmonisation existantes ne pouvaient pas être adaptées à un usage clinique en raison de limitations intrinsèques. Par exemple, la plupart d’entre elles ne peuvent pas généraliser leur apprentissage à des données non vues. Par conséquent, nous avons proposé un modèle original d’apprentissage profond (deep-learning) d’harmonisation, ‘ImUnity’. Adoptant une approche 2.5D, il dérive d’une architecture de type VAE-GAN à laquelle se greffent un module biologique et de ‘désapprentissage’. Cette approche a été évaluée à l’aide de 3 bases de données open source (ABIDE, OASIS et SRPBS), contenant des IRM provenant de plusieurs types de scanners d’acquisition ou de fournisseurs, avec des sujets d’une large tranche d’âge.

Par la suite, nous avons évalué l’impact d’ImUnity sur le développement des volumes cérébraux et des épaisseurs corticales, en utilisant la base de données ABIDE. ImUnity a permis

de rapprocher l'évolution du volume cérébral avec l'âge, avec des tendances de la littérature basées sur de grandes études mono-centriques. Cela fut moins évident pour l'épaisseur corticale, car elle semble moins concernée par les problèmes d'harmonisation. Des investigations plus approfondies doivent être menées pour compléter cette dernière étude sur l'impact de l'harmonisation sur le développement du cerveau.

Keywords— Développement cérébral, Deep Adversarial Network, harmonisation de données

CONTENTS

7.1	Introduction	98
7.2	Résumé de chapitre : Bibliographie	100
7.3	Résumé de chapitre : Évaluation de ComBAT et cycleGAN pour l'harmonisation de 2 sites	102
7.4	Résumé de chapitre : ImUnity, un VAE-GAN adapté à l'har- monisation de données IRM multi-centrique	105
7.5	Résumé de chapitre : Proposition d'évaluations des effets de l'harmonisation sur le développement cérébrale chez l'enfant . .	107
7.6	Conclusion et perspectives	110

7.1 Introduction

Ce projet de thèse a émergé du projet national français DEFIDIAG (Binquet et al., 2022). Il est soutenu par MIAI, le 'Multidisciplinary Institute in Artificial intelligence' de Grenoble, et fait partie de la chaire multiomique.

DEFIDIAG, pour 'Intellectual Disability Diagnostic', est un projet pilote financé par le programme "Plan France Médecine Génomique 2025 PFMG 2025" portant sur l'étude de la déficience intellectuelle (DI). Parmi les maladies rares, la DI est la première cause de consultation dans les centres de génétique pédiatrique et on estime que 1 à 3% de la population est atteinte de DI. Elle est extrêmement difficile à diagnostiquer car des centaines de gènes sont impliqués dans ces pathologies, rendant difficile l'identification des causes de la maladie et l'adaptation des soins aux patients. DEFIDIAG vise à "démontrer la faisabilité du séquençage complet du génome, ainsi que son efficacité, en première intention, dans la détermination des gènes impliqués dans la déficience intellectuelle". Au total, 1 275 patients ainsi que leurs deux parents biologiques devraient participer à l'étude sur une durée de plus de 3 ans. Cela "permettra à davantage de familles de connaître plus rapidement les causes de la maladie, de bénéficier dans un délai plus court d'une prise en charge adaptée et peut-être de prévenir la survenue de complications spécifiques" (<https://defidiag.inserm.fr/accueil/the-defidiag-program>).

Intrinsèquement, DEFIDIAG se concentre sur les analyses génétiques. L'étude recueille également des données d'imagerie cérébrale par résonance magnétique sur des sujets répartis sur 11 sites dans toute la France. Grâce à l'expertise de notre laboratoire en matière de traitement d'images médicales, notre projet de thèse vise à combiner ces données radiomiques avec l'analyse génétique afin d'acquérir des connaissances sur ces maladies rares et d'améliorer leur diagnostic.

A ce jour, les données DEFIDIAG sont toujours en cours de collecte. Par conséquent, nos travaux ont été consacrés au développement d'outils de traitement d'images adaptés aux spécificités des données DEFIDIAG et permettant d'ouvrir la voie aux analyses futures. En particulier, 2 préoccupations majeures ont retenu notre attention : (1) les sujets de la base de données sont principalement des enfants, et l'effet de l'âge sur le développement du cerveau au cours de cette tranche d'âge (5-25 ans) doit être pris en compte afin d'éviter les facteurs de confusion. (2) Dans le contexte de DEFIDIAG, et en général pour toutes les études portant sur des maladies rares, le nombre de patients peut être très faible et peut limiter la fiabilité des résultats. Par conséquent, il est courant de rassembler des données provenant de différents centres d'acquisition. Cependant, au cours du programme DEFIDIAG, ces acquisitions n'ont pas été réalisées en suivant un protocole d'imagerie spécifique à l'étude. Chaque site d'inclusion a suivi ses propres procédures d'imagerie. Nous attendons donc de grandes variations dans l'ensemble de données en raison des différents scanners et éventuellement des différents protocoles d'acquisition. Le nombre de patients par site étant limité (quelques dizaines voire quelques centaines), ces fluctuations induites par les sites peuvent avoir des effets plus importants que les variations biologiques d'intérêt. Il est donc primordial de considérer ces variations et d'essayer de les supprimer. Les méthodes permettant de corriger ces effets multi-centriques sont appelées 'méthodes d'harmonisation' et sont devenues notre principal point d'intérêt durant cette thèse.

Compte tenu du contexte énoncé, cette thèse s'est focalisée sur le développement de nouveaux outils de Deep-Learning (DL) pour l'harmonisation d'IRM cérébrales. Les données

DEFIDIAG n'étant pas disponibles, il a été choisi de tester nos solutions sur des bases de données d'imagerie en accès libre, présentant des caractéristiques aussi proches que possible de celles de DEFIDIAG. La base de données ABIDE [Di Martino et al. \(2014\)](#), qui est une importante base multi-centrique axée sur les Troubles du Spectre Autistique (TSA) chez l'enfant, a finalement été choisie comme référence. Toutes les études réalisées au cours de cette thèse ont été effectuées sur cette base ABIDE. Deux autres bases de données ont été utilisées à des fins de validation car elles présentent des acquisitions de 'traveling subjects' (sujets scannés sur différents sites) et permettent une évaluation quantitative.

- **REVUE BIBLIOGRAPHIQUE**

La première partie de cette thèse a consisté à rassembler les résultats de la littérature en rapport avec notre projet. Elle a commencé par une revue de la littérature portant sur le développement normal du cerveau durant l'enfance (voir [sous-section 2.1.1](#)). Cette étude fut essentielle pour les recherches ultérieures, qui ont été effectuées sur des données d'enfants afin de correspondre au contexte initial. En plus des sujets sains, nous nous sommes ensuite intéressés aux troubles cérébraux neurologiques et à leurs impacts sur le développement cérébral. Enfin, nous avons examiné l'impact du développement du cerveau pour son recalage sur un atlas de référence, une étape de pré-processing classique en imagerie médicale. En raison des grandes variations cérébrales attendues au cours de l'enfance et de l'adolescence, des atlas adaptés à chaque tranche d'âge ont été sélectionnés.

Notre deuxième analyse bibliographique a porté sur l'harmonisation de données (voir [section 2.2](#)). C'est un sujet qui fut introduit pour la première fois en 1998, suivi par de nombreuses avancées depuis. Les solutions statistiques basées sur l'algorithme de correspondance des histogrammes ont été les premières à donner des résultats prometteurs et beaucoup de solutions s'en sont inspirées. Par la suite, des solutions inspirées du monde de la génomique, où le 'batch effect' doit être supprimé, ont été proposées. ComBAT est l'une d'entre elles, et est toujours considérée aujourd'hui comme une référence en harmonisation de données IRM. Enfin, sous l'influence de la popularité croissante du DL, des modèles initialement développés pour la segmentation ou la génération d'images ont été adaptés à l'harmonisation. La plupart des études récentes sont basées sur le DL, utilisant les architectures U-Net, cycleGAN ou VAE, et ont montré des résultats très prometteurs.

- **COMPARAISON DES MÉTHODES DE L'ÉTAT DE L'ART**

Notre première étude ([chapitre 3](#)) a comparé deux solutions d'harmonisation de la littérature. L'idée était de comparer un modèle DL récent (cycleGAN) à une solution linéaire de référence (ComBAT) afin d'estimer le potentiel des solutions d'harmonisation basées sur DL. Cela a aussi permis de proposer un workflow original pour évaluer les résultats de l'harmonisation. Dans notre étude, nous nous sommes concentrés sur la suppression des bruits induits par le site mais aussi sur l'impact de l'harmonisation sur des bruits synthétiques que nous avons artificiellement ajoutés aux images. Nous avons considéré deux types de bruits : (1) les 'bruits globaux' qui imitent les bruits induits par le scanner et que nous souhaitons supprimer, et (2) les 'bruits locaux' pouvant être assimilés à des variations locales biologiques (traumatisme, accident vasculaire cérébral, tumeur...), devant être préservés après l'harmonisation. L'évaluation s'est basée sur les 'données radiomiques' extraites des images avant et après harmonisation. Un Support Vector Machine (SVM) a ensuite été utilisé pour détecter la présence ou l'absence du

signal d'intérêt (bruit induit par le site, bruit synthétique global ou local).

- **PROPOSITION D'UN MODÈLE ORIGINAL D'HARMONISATION DL**

Suivant notre première étude qui a montré que les techniques DL peuvent être adaptées à l'harmonisation de données IRM, nous avons constaté dans la revue de la littérature que chaque solution DL présente des conditions spécifiques d'utilisation, dues à leur architecture. Par exemple, cycleGAN est limité à l'harmonisation de données de deux sites seulement. Cette condition est rarement satisfaite en pratique rendant ce modèle non adapté aux grandes bases de données multi-centriques. D'autres exigences peu réalistes concernent 1) le besoin de 'traveling subjects' pour entraîner la solution ; 2) le besoin d'ajuster la solution à chaque fois que des données provenant d'un nouveau site ou d'un nouveau scanner sont ajoutées ; 3) le besoin de ré-harmoniser les données pour chaque application clinique. Dans [chapitre 4](#), nous avons proposé une approche originale d'harmonisation qui ne repose pas sur ces conditions et est donc adaptée à la pratique clinique. Cette nouvelle solution s'appelle ImUnity et est basée sur une architecture VAE-GAN. L'étape de validation de cette méthode a représenté la principale innovation proposée dans cette thèse. Cette étude rassemble différentes expériences portant sur l'effet de suppression de site, la prédiction du diagnostic de TSA et l'amélioration de la similarité en utilisant des 'traveling subjects'.

- **EFFETS DE L'HARMONISATION SUR LE DÉVELOPPEMENT CÉRÉBRAL**

Afin d'approfondir notre processus de validation, nous avons finalement testé ImUnity sur une application clinique liée au programme DEFIDIAG. L'étude est présentée dans le [chapitre 5](#). Nous avons évalué l'effet de l'harmonisation sur l'évolution du volume apparent du cerveau et de l'épaisseur corticale au cours de l'enfance. Cette évaluation a été réalisée sur des sujets sains et des patients atteints de TSA, combinant les 11 sites d'acquisition de ABIDE. Les évolutions des mesures de volumes et d'épaisseur corticales avec l'âge, mesurées avant et après harmonisation, ont été comparées à celles trouvées dans la littérature dans de grandes études mono-centriques. Une partie de l'étude a été menée en collaboration avec Constance Sohler, une étudiante ingénieur qui a effectué son stage de fin d'études sous ma supervision. Notre travail suggère un impact positif d'ImUnity, qui a réduit l'effet de site tout en améliorant les scores biologiques. D'autres analyses statistiques doivent encore être effectuées afin de conclure correctement cette étude.

7.2 Résumé de chapitre : Bibliographie

Le projet DEFIDIAG, et plus généralement les études impliquant des images du cerveau d'enfants, présentent des défis spécifiques. Au cours du développement du cerveau, les volumes cérébraux changent globalement, mais aussi localement dans les régions corticales et sous-corticales. Afin d'être en mesure de traiter les données correctement et de mettre en commun des données multiples provenant de différentes tranches d'âge, nous avons commencé par rechercher des références sur le développement cérébral, puis sur les outils appropriés pour le pré-processing de ces images.

Nous nous sommes premièrement concentrés sur l'évolution cérébrale (volume et épaisseur corticales) avec l'âge. Nous avons rassemblé les résultats d'études mono-centriques et longitudinales de la littérature. Ces études s'intéressent à l'évolution du cerveau chez des sujets sains mais aussi chez des patients atteints de TSA, ce qui a permis de mieux comprendre les

changements structurels induits par ces troubles intellectuels. Ces résultats ont été utilisés dans le dernier chapitre ([chapitre 5](#)) comme références pour évaluer l'effet de l'harmonisation des images.

Le second sujet documenté est l'étape de "recalage", qui consiste à aligner les scans sur un cerveau de référence spécifique, également appelé "template". Ce template représente un cerveau humain moyen en bonne santé. Cependant, des études ont montré que l'utilisation de templates classiques (type 'MNI') pour des études traitant des données d'enfants peut entraîner une perte d'informations. Ceci est une conséquence directe de la morphologie du cerveau et des variations structurelles avec l'âge, que ce soit globalement ou localement [sous-section 2.1.1](#). Pour pallier ce problème, [Sanchez et al. \(2012a\)](#) ont proposé une base de données en libre accès contenant des templates à intervalles de 6 mois pour la tranche d'âge de 4,5 à 19,5 ans. Dans cette thèse, nous avons utilisé cette base de données pour l'étape de recalage dans toutes les expériences impliquant des données d'enfants.

Notre revue de la littérature s'est ensuite intéressée aux principaux défis de l'harmonisation, et comment le domaine est passé de solutions statistiques classiques à des solutions plus complexes avec la popularisation du DL. [section 2.2](#) présente en détails les solutions actuelles afin de mieux comprendre leurs principales forces et faiblesses.

Nous présentons dans un premier temps les solutions statistiques adaptées à l'harmonisation existantes ([sous-section 2.2.1](#)). Celles-ci sont principalement inspirées de l'algorithme d'alignement d'histogrammes ([section 2.2.1](#)) visant à aligner les distributions d'intensité des sites considérés. Ces méthodes ont montré de bonnes fonctionnalités en termes de temps de calcul et d'efficacité. Cependant, elles ont le défaut de supprimer certaines variations locales d'intensité, ce qui pénalise fortement les analyses futures dans le cas où celles-ci sont d'origine biologique.

Les principales solutions qui ont succédé aux méthodes d'alignement d'histogrammes ont été le fruit de travaux d'une équipe du département de bio-statistiques de l'Université de Pennsylvanie. Ils ont proposé 3 solutions importantes, à savoir le 'White-Stripe normalization', 'RAVEL' et l'adaptation de ComBAT à des données IRM. Cette dernière solution a particulièrement fait parler d'elle et est encore aujourd'hui considérée comme référence dans le domaine de l'harmonisation. Lors d'une étude comparative, [Fortin et al. \(2017\)](#) ont montré que ComBAT était plus performant pour modéliser et supprimer la variabilité indésirable entre les sites dans les cartes FA et MD que les autres méthodes existantes. ComBAT semble préserver à la fois la variabilité biologique et supprimer les variations indésirables introduites par le site beaucoup mieux que les autres techniques. Depuis le papier original en 2017, plusieurs études proposant des versions modifiées de ComBAT ont été publiées ([Beer et al., 2020](#); [Da-ano et al., 2020](#); [Pomponio et al., 2020](#)).

Au cours des dix dernières années, l'apprentissage profond a été largement utilisé dans les études d'imagerie médicale en général. Ce fut également le cas pour l'harmonisation de données IRM, qui peut être considérée comme un problème de 'domain adaptation' (DA). Nous nous sommes donc intéressés aux solutions de DL existantes pour l'harmonisation d'images médicales. Ces modèles génératifs visent à supprimer le bruit propre aux sites en générant des images "corrigées". Une fois réalisé, le scan harmonisé peut être utilisé dans n'importe quel processus clinique. Pour être efficaces, ces solutions doivent être adaptées à n'importe quel site ou scanner et ne doivent pas nécessiter d'informations supplémentaires par rapport au scan à harmoniser. Dans le cas contraire, leur intégration dans le pipeline clinique

serait compromise. De nombreuses nouvelles techniques génératives ont été proposées, et sont pour la plupart basées sur le principe du Generative Adversarial Network (GAN ; [Goodfellow et al. \(2014\)](#)). Les modèles présentés en détails ([sous-section 2.2.2](#)) sont ‘DeepHarmony’ ([Dewey et al., 2019](#)), ‘CycleGAN’ [Zhu et al. \(2018\)](#), ‘Calamity’ ([Zuo et al., 2021b](#)) et StarGAN ([Choi et al., 2018](#)).

Inspirées par les techniques de DA, certaines équipes ont proposé d’aborder l’harmonisation directement dans leur modèle utilisé pour leur application clinique. Au lieu de générer des images harmonisées, ces solutions proposent d’estimer les bruits indésirables tout en abordant une application clinique (segmentation, classification, etc...). Elles consistent généralement à intégrer un module de ‘désapprentissage’ au modèle en le connectant à l’espace latent. L’élagage (‘Pruning’) est une autre technique de ‘désapprentissage’ et consiste à supprimer les connexions neuronales à l’intérieur du modèle pour l’empêcher l’overfitting ou d’apprendre des caractéristiques liées aux différents sites.

Ces solutions semblent très intéressantes car elles ne génèrent pas de nouvelles images. Cette étape nécessite généralement beaucoup de données et les modèles peuvent avoir du mal à converger (voir [sous-section 2.2.2](#)). D’autre part, comme ces modèles ne génèrent pas d’images harmonisées, des modules de ‘désapprentissage’ doivent être introduits et entraînés pour chaque application clinique. À l’inverse, les modèles génératifs d’harmonisation cherchent à produire des images qui peuvent être utilisées pour toute application clinique. [sous-section 2.2.3](#) présente en détails les solutions existantes. A savoir ‘Attention-guided deep domain adaptation’ ([Guan et al., 2021](#)), ‘Unlearning module’ [Dinsdale et al. \(2021\)](#), et ‘model pruning’ [Dinsdale et al. \(2022\)](#).

Enfin, nous avons proposé un tableau comparatif ([2.1](#)) des méthodes DL qui résume leurs différentes caractéristiques.

7.3 Résumé de chapitre : Évaluation de ComBAT et cycleGAN pour l’harmonisation de 2 sites

Dans cette étude, nous avons comparé ComBat et cycleGAN dans le cadre d’harmonisation de données T1w. Nous avons réalisé cinq expériences différentes sur des données synthétiques ainsi que sur des images réelles *in vivo*. Nous avons d’abord évalué la capacité des 2 méthodes à supprimer des bruits globaux ajoutés manuellement dans les images ainsi que leur capacité à préserver des variations locales ajoutées manuellement. Nous avons également étudié l’impact de ces méthodes sur des tâches de classification de sites et la détection des Troubles du Syndrome Autistique (TSA). Les effets de l’harmonisation ont été évalués en utilisant des métriques radiomiques, connues pour être sensibles à l’harmonisation ([Da-ano et al., 2020](#); [Orlhac et al., 2019](#)). Enfin, nous avons évalué l’impact des solutions d’harmonisation sur différents groupes de données radiomiques. Ces métriques sont censées représenter les principales caractéristiques des images, comme la forme, le contraste ou la texture.

Pour comparer ces 2 méthodes, nous avons réalisé 3 expériences sur des données synthétiques et 2 sur des données réelles *in vivo*. Les données synthétiques ont été utilisées pour évaluer la capacité à supprimer le bruit global ou à préserver les structures locales connues ou les variations biologiques locales. D’autre part, les données réelles ont été utilisées pour estimer l’efficacité des méthodes à supprimer les effets de site, et leur capacité à améliorer les analyses cliniques ultérieures. CycleGAN a été entraîné à partir de zéro pour chaque ex-

périence. Pour toutes les tâches de classification, nous avons utilisé un SVM pour classifier les données avant et après harmonisation. Pour évaluer la spécificité et la sensibilité de notre classificateur, nous avons utilisé l'aire sous la courbe (AUC) de la courbe ROC (Receiver Operating Characteristic). L'inspection visuelle n'étant pas suffisante pour évaluer l'effet de l'harmonisation sur les images, nous avons extrait des métriques radiomiques. Nous avons utilisé l'API python pyradiomics développée par [van Griethuysen et al. \(2017\)](#) pour extraire 101 métriques. Dans tous les cas, nous avons d'abord sélectionné les "métriques les plus corrélées" à l'aide de tests de Pearson (exécutés indépendamment pour chaque métrique) avec la métriques d'intérêt (affiliation au site, présence de bruit ajouté, etc...). Nous avons également effectué des tests de Pearson après harmonisation sur des métriques radiomiques préalablement sélectionnées afin de mieux comprendre l'impact des deux méthodes sur ces métriques. Enfin, nous avons étudié la corrélation entre les métriques radiomiques et les variables biologiques (sexe et âge). Notre hypothèse était que l'harmonisation devait augmenter ou au moins préserver la corrélation lorsqu'elle existait. Toutes les procédures de classification ont été effectuées en utilisant une '8-fold cross validation' (tandis qu'une '10-fold cross validation' a été utilisée pour entraîner notre modèle cycleGAN) répétée 10 fois afin d'avoir différentes combinaisons d'ensembles et de pouvoir calculer des évaluations statistiques sur la performance des classifieurs. Pour visualiser les résultats, nous avons réduit la dimension avec PCA et TSNE : ([Maaten and Hinton, 2008](#)). La PCA a d'abord été utilisée pour assurer une représentation orthogonale de nos données (8 composantes utilisées, représentant environ 95% de la variance totale), puis TSNE pour représenter visuellement nos données selon 2 axes. Une fois les dimensions réduites, il a été possible de visualiser deux groupes de points correspondant à des sites ou des types de données différents. Pour la validation, nous avons uniquement utilisé les données réduites par la PCA, en utilisant les 8 premières composantes principales.

Enfin, des 'Welch t-tests' ([Welch, 1947](#)) ont été exécutés pour valider si les résultats étaient statistiquement significatifs ou non. Nous avons effectué ces tests sur toutes les combinaisons de données sous les hypothèses nulles "la méthode n'a pas d'impact sur les performances du SVM" et "les deux méthodes ont les mêmes performances". Nous avons ensuite observé les p-valeurs de ces tests et rejeté l'hypothèse H_0 si $p < 0.05$. Comme les variances des résultats obtenues par les deux méthodes ne pouvaient pas être considérées comme égales, nous avons utilisé le 'Welch t-test' pour comparer si les différences observées étaient statistiquement significatives.

Pour la première expérience qui visait à évaluer la capacité des deux méthodes à supprimer les bruits globaux ajoutés manuellement aux images, nous avons ajouté une variation Gaussienne globale d'intensité centrée au milieu des images. Ensuite, un bruit Gaussien classique a été ajouté pour induire des artefacts multiples et réduire le contraste des images.

Pour vérifier que les variations locales des intensités étaient conservées après l'harmonisation, une variation Gaussienne d'intensité a été ajoutée localement à certaines données échantillonnées de manière aléatoire. Cela imite les hyper-signaux que l'on peut trouver chez les patients présentant plusieurs pathologies (AD, les gliomes ou les tumeurs) ou chez les sujets ayant subi un accident vasculaire cérébral ou un traumatisme. Pour cette expérience, cycleGAN et ComBAT ont été entraînés sur des données saines non modifiées, puis utilisés sur l'ensemble du jeu de données (données originales + données modifiées).

La quatrième expérience évaluait l'harmonisation de 2 sites sélectionnés via la capacité d'un SVM à classifier l'origine de chaque image. L'hypothèse ici était qu'aucun classificateur ne devrait être capable de détecter avec précision l'origine des données. Nous avons donc

essayé de classer l’origine des données avant et après l’harmonisation pour voir comment l’harmonisation pouvait pénaliser le classificateur.

Dans notre dernière expérience, nous avons exécuté une tâche de classification clinique (patients atteints de TSA vs. sujets sains) sur les données des sites A et B. Cette classification a été effectuée avant et après harmonisation afin d’évaluer si les SVM sont plus performants sur des données harmonisées que sur des données brutes. L’hypothèse ici était que l’harmonisation devait préserver l’information biologique. Ainsi, nous nous attendions à obtenir de meilleures métriques de classification après harmonisation.

Les résultats obtenus ont confirmé la nécessité d’harmoniser les données et ont montré l’efficacité de ComBAT et de cycleGAN pour l’harmonisation de données lors d’études multi-centriques. Cette étude a montré que les deux méthodes réduisaient les bruits globaux ajoutés et les effets de site, tout en conservant les modifications locales, et en améliorant la précision du SVM pour la classification des lésions synthétiques et des patients atteints de TSA. Cependant, il est important de souligner les différences entre les performances des deux méthodes. Alors que ComBAT semble être plus adapté pour supprimer les bruits globaux et améliorer la corrélation entre les métriques radiomiques et l’affiliation au site ou l’âge, cycleGAN montre de meilleurs résultats pour préserver les modifications locales et améliorer l’analyse statistique des études cliniques.

Les résultats concernant les différences entre les nombres de caractéristiques significativement corrélées dans [tableau 3.4](#) sont très intéressants. Cette différence peut s’expliquer par le fait que l’algorithme ComBAT est conçu pour éliminer les effets de l’affiliation au site tout en préservant les corrélations avec l’âge (puisqu’il prend l’âge et l’affiliation au site comme entrées). A l’inverse, cycleGAN ne prend que les images originales en entrée. Il pourrait donc être intéressant d’ajouter d’autres entrées biologiques comme l’âge et le sexe au réseau pour voir comment cela pourrait affecter les résultats de [tableau 3.4](#). En examinant les résultats des tests de Pearson sur les métriques radiomiques, nous avons constaté que les deux méthodes préservent les variables liées à la forme, comme prévu. Le contraire aurait en effet été problématique, car les bruits liés au site n’altèrent pas l’anatomie des images mais impactent leur contraste. Les impacts des deux méthodes sur les autres familles de métriques se sont révélés complémentaires : ComBAT a obtenu de bons résultats sur les caractéristiques GLRLM alors que les caractéristiques GLSZM ont mieux bénéficié de l’harmonisation de cycleGAN. Les autres familles ont été affectées de manière similaire par les deux algorithmes. Il serait donc intéressant d’étudier des combinaisons des deux méthodes.

Pour illustrer l’adaptabilité des 2 méthodes à un faible jeu de données, nous avons également réalisé l’expérience 4 ([section 3.2.4](#)) sur les sites A et B en utilisant 20 sujets témoins uniquement. Nous avons constaté que cycleGAN était limité par la taille de l’échantillon et n’était pas en mesure de corriger les effets de site, tandis que ComBAT a présenté des résultats similaires à ceux obtenus avec l’ensemble des données.

Pour conclure, nous avons montré durant cette étude l’importance de l’harmonisation lorsqu’il s’agit de données provenant d’au moins 2 centres. En effet, nous avons pu distinguer avec précision les données provenant de deux sites d’acquisition, bien qu’ils aient utilisé le même type de scanners. Les deux approches ont permis de supprimer efficacement les effets de site indésirables tout en préservant les informations biologiques. Elles ont montré un impact positif dans toutes les expériences étudiées, comme prévu. Elles ont de plus permis d’améliorer les mesures de classification des patients atteints de TSA. Dans ce cas précis,

CycleGAN a donné de meilleurs résultats que ComBAT, tandis que ce dernier a pu mieux réduire les bruits globaux dans les images. Les résultats de CycleGAN ont démontré que les solutions de DL semblent adaptées à l’harmonisation et pourraient surpasser les méthodes statistiques classiques (linéaires) telles que ComBAT dans certaines conditions. De manière surprenante, nous avons également montré que cycleGAN n’étaient pas toujours meilleur que ComBAT. Les deux méthodes apparaissent complémentaires sur plusieurs aspects et n’ont pas les mêmes effets sur les familles radiomiques. Cela pourrait déterminer le choix des techniques en fonction de l’objectif à atteindre. En outre, cela ouvre la voie à de nouvelles solutions capables de tirer parti des deux méthodes présentées. De plus, le fait que ces deux solutions ne puissent pas être utilisées sur de nouveaux sites souligne la nécessité d’une solution plus adaptée. Sinon, il est très peu probable qu’une solution d’harmonisation soit utilisée dans la pratique clinique.

7.4 Résumé de chapitre : ImUnity, un VAE-GAN adapté à l’harmonisation de données IRM multi-centrique

Inspirés par les dernières avancées dans les solutions d’harmonisation générative (Dewey et al., 2020; Zuo et al., 2021a) et de ‘désapprentissage’ (Dinsdale et al., 2020; Guan et al., 2021), nous avons proposé un modèle original d’harmonisation, ImUnity. Cette solution est basée sur un modèle de DL 2.5D et permet une harmonisation rapide et flexible. ImUnity génère des images RM "corrigées" qui peuvent ensuite être utilisées pour diverses études d’imagerie. Ce modèle, basé sur une architecture VAE-GAN, utilise pour son apprentissage plusieurs coupes provenant du même individu et des transformations de contraste d’image aléatoires. Il désapprend également l’information liée aux sites à l’aide d’un module de confusion connecté à l’espace latent. De façon similaire, un module biologique assure la préservation de l’information clinique (sexe, âge, pathologies...). Une fois entraînée, cette architecture doit permettre d’harmoniser les données provenant de nouveaux sites ou scanners sans qu’il soit nécessaire de ré-entraîner le modèle. Il est aussi possible d’harmoniser les images vers plusieurs ‘contrastes de référence’. L’utilisateur peut alors choisir arbitrairement plusieurs reconstructions d’images RM en fonction du contraste choisi. L’architecture de ImUnity est représentée en [figure 4.1](#). Le code de ImUnity est en accès libre [ici](#).

Notre modèle 2.5D combine les prédictions de 3 modèles 2D, chacun entraîné le long d’un axe spécifique. Cette approche a été introduite pour la première fois dans le domaine de l’harmonisation par (Dewey et al., 2019), soulignant le grand potentiel d’une telle approche. Comme l’illustre [figure 4.4](#), cela permet d’obtenir un résultat final de meilleure qualité sans discontinuité suivant un axe comme il est fréquemment le cas avec des modèles 2D.

Pour évaluer l’efficacité et la flexibilité de notre outil d’harmonisation, nous avons évalué l’approche en utilisant 3 bases de données open source contenant des images provenant de plusieurs sites d’acquisition, scanners ou d’intensité de champs magnétiques, et d’un large éventail d’âges de patients. Pour la plupart des expériences, ImUnity a été entraîné en utilisant les données d’une seule des bases de données, puis appliqué aux deux autres pour évaluer la capacité de généralisation du modèle. La qualité des images reconstruites, la capacité à éliminer les biais liés au site ou au scanner et la capacité à classer les patients ont été évaluées après harmonisation des données.

Les données extraites des trois bases ont été utilisées pour évaluer différents aspects de

notre modèle. L’impact sur la qualité de l’image dans le cadre d’une harmonisation multi-sites ou multi-scanner a été évalué à l’aide de données provenant de ‘traveling subjects’ (permettant d’avoir une vérité terrain) issues des ensembles de données OASIS et SRPBS. La capacité à supprimer les informations liées au site a été évaluée en utilisant les données ABIDE. Enfin, la capacité du modèle concernant l’harmonisation inter-sites a été évaluée en utilisant la prédiction des troubles autistiques chez les enfants de la base ABIDE. Pour démontrer la flexibilité d’ImUnity, toutes les expériences ont été réalisées avec le même modèle, entraîné sur des données provenant de la base ABIDE, sauf indication contraire. OASIS et SRPBS ont été utilisés uniquement pour les parties de validation. Chaque modèle a été entraîné sur des coupes 2D comportant au moins 1% de voxels de tissu cérébral. L’apprentissage a été effectué sur une Nvidia GeForce 2080 RTX pendant 300 époques en utilisant une ‘learning rate’ de 10^{-4} ainsi que l’optimiseur Adam.

Les résultats obtenus ont montré des performances élevées en termes de qualité d’images harmonisées, ainsi qu’une élimination claire du biais lié aux conditions d’acquisition des images. De plus, les expériences réalisées ont clairement démontré la polyvalence d’ImUnity. En entraînant le modèle sur des ensembles de données extraits d’une base de données (ici ABIDE) et en s’intéressant à l’harmonisation de ‘traveling subjects’ fournis par deux bases de données différentes (ici OASIS et SRPBS), nous avons montré qu’ImUnity ne nécessite pas de nouvelle phase d’entraînement pour s’adapter à des sites ou des scanners non vus (voir [figure 4.3](#)). De plus, les performances ont été maintenues indépendamment du site choisi comme une référence (voir [tableau 4.2](#)). Bien que le modèle ait été entraîné sur les données ABIDE uniquement, il a démontré de meilleurs résultats que la littérature en termes de qualité d’image (+4%, voir [tableau 4.2](#)).

Le [tableau 4.2](#) présente les mesures SSIM obtenues lors de l’entraînement de ImUnity sur les deux autres ensembles de données (OASIS et SRPBS). Comme aucune caractéristique biologique n’était disponible dans ces bases de données, le module biologique a été désactivé et le modèle a été entraîné de manière auto-supervisée. Premièrement, nous avons noté de meilleurs résultats concernant l’harmonisation multi-scanners lorsque le modèle a été entraîné et appliqué sur la même base de données (ici OASIS, +2,5%). Deuxièmement, le score obtenu pour l’harmonisation multi-sites (SRPBS) était le plus élevé lorsque le modèle était entraîné sur les données OASIS ($N = 1098$) (avec un score légèrement meilleur qu’avec les autres bases de données). Nous avons aussi pu observer l’impact sur ces scores de la taille du jeu de données, du nombre de sites/scanner impliqués dans la formation, de l’utilisation du module biologique et des différences anatomiques entre les jeux de données (ABIDE contenant principalement des données d’enfants alors que OASIS et SRPBS se concentrent sur les adultes). Alors que les résultats d’OASIS suggèrent une meilleure généralisation sur des données non vues parce que plus de données d’entraînement étaient disponibles, les résultats d’ABIDE suggèrent que les différences anatomiques pourraient être compensées par un grand ensemble de données d’entraînement présentant plus de variabilité site/scanner que OASIS (11 sites pour ABIDE contre 4 sites pour OASIS). D’autre part, les résultats de l’harmonisation multi-scanners ont montré la difficulté du modèle entraîné sur les données SRPBS à généraliser son entraînement aux données OASIS. Cela indique un effet d’‘overfitting’ dans cette situation, car il n’y avait pas assez de données d’entraînement (ici $N=81$ réparties sur 9 sites). Cela suggère que ImUnity n’est peut-être pas adapté aux scénarios de petite taille d’échantillon, ce qui fournit des informations utiles pour comprendre pourquoi le modèle est moins efficace dans certains contextes. L’augmentation des données aurait pu aussi être

ajoutée pour améliorer les résultats de cette expérience.

Pour mieux estimer l'impact du module biologique, nous avons également exécuté l'expérience 3 (section 4.2.4) sans celui-ci. Nous avons constaté (voir figure 4.8) que le module biologique a un impact positif sur les résultats avec une contribution représentant environ 20% de l'effet d'harmonisation total et suggérant que des informations supplémentaires en entrée en plus de l'image pourraient conduire à de meilleurs résultats en sortie. De plus, dans l'expérience 1, nous avons testé différentes combinaisons de données de formation et de test. Dans la situation : 'training = OASIS' et 'testing = SRPBS', aucune caractéristique biologique n'était disponible et le module biologique n'a donc pas été utilisé. Notre solution a tout de même montré une bonne généralisation sur l'ensemble de données de test. Sur la base de cet ensemble de résultats, nous pouvons supposer que l'ajout de certaines informations biologiques au modèle conduirait également à de meilleurs résultats sur les 'traveling subjects'.

En conclusion, ce chapitre présente ImUnity qui est un outil original et efficace dédié à l'harmonisation de données IRM. Le modèle 2.5D que nous proposons dérive de l'architecture VAE-GAN. Il garantit des résultats réalistes et permet de supprimer les biais liés aux sites et de préserver les informations biologiques. Nos résultats montrent que la méthode atteint les performances de la littérature en termes de qualité d'image sur les 'traveling subjects' des bases de données OASIS et SRPBS. ImUnity a de plus permis d'améliorer la classification des patients autistes de la base de données ABIDE. Le modèle ne nécessite qu'un seul type de séquence IRM sans qu'il soit nécessaire de faire correspondre les sujets, il peut être généralisé à des sites non vus pendant la phase d'entraînement et peut être utilisé pour harmoniser les images IRM vers différents contrastes de référence sans nouvelle phase d'entraînement.

Des perspectives d'amélioration subsistent. Comme l'introduction de variations de contraste plus complexes durant l'entraînement et l'utilisation de techniques de DA plus spécifiques pourraient permettre l'harmonisation de plusieurs séquences d'IRM à la fois. Même si la solution proposée a été validée sur une expérience clinique et s'est avérée efficace, elle doit encore être testée de manière plus générale. Plusieurs études d'harmonisation (Beer et al., 2020; Fortin et al., 2018) ont étudié l'impact de l'harmonisation sur le développement du cerveau, sur lequel nous disposons de résultats solides issus de grandes études mono-centriques comme présentées dans sous-section 2.1.1.

7.5 Résumé de chapitre : Proposition d'évaluations des effets de l'harmonisation sur le développement cérébrale chez l'enfant

Dans ce dernier chapitre, nous avons évalué la qualité des images harmonisées via une étude sur le développement du cerveau entre de 5 à 25 ans. Les sujets considérés sont issus de la base de données ABIDE. chapitre 3 et chapitre 4 ont témoigné d'un important besoin d'harmonisation sur la base de données ABIDE. Ces variations indésirables impactent très probablement le développement biologique observé du cerveau, modifiant en conséquence les tendances d'évolution des ROIs avec l'âge. Ici, les données provenant de 11 sites ont été regroupées et analysées avant et après harmonisation. Pour chaque sujet, les volumes et l'épaisseur des régions corticales et sous-corticales d'intérêt ont été extraits via l'outil

FreeSurfer. À des fins de comparaison, nous avons également utilisé ComBAT ([section 2.2.1](#)).

Nous avons utilisé 271 sujets sains et 253 patients atteints de TSA âgés de 5 à 25 ans. Partant du principe que l’harmonisation des données aiderait à retrouver les trajectoires de développement attendues, nous nous sommes référés à plusieurs tendances de la littérature déjà présentées au [chapitre 2](#) ([Ducharme et al., 2016](#); [Lenroot and Giedd, 2006](#); [Vijayakumar et al., 2016](#); [Wierenga et al., 2014](#)), ce qui nous a permis d’évaluer les effets de ImUnity et ComBat. Ces tendances ont été recueillies à partir de grandes études mono-centriques, évitant tout problème lié à l’harmonisation. L’effet de l’harmonisation a été défini par son impact sur les tendances d’évolution du volume, et la façon dont l’harmonisation les a rapprochées de celles rapportées dans la littérature. Pour cette expérience, nous avons considéré séparément les sujets sains et les sujets atteints de TSA car cette pathologie a un impact significatif sur le développement des régions corticales et sous-corticales du cerveau pendant la petite enfance (voir [sous-section 2.1.2](#)). Dans cette étude, nous n’avons pas pu inclure les femmes atteintes de TSA car nous n’avons pas pu recueillir suffisamment de résultats de la littérature. De même, pour l’évolution de l’épaisseur corticale des ROIs du cerveau, seuls les hommes (sains et ASD) ont pu être considérés.

Cette étude s’est déroulée en 4 expériences. Une première évaluation visuelle des tendances a été effectuée afin d’observer l’impact des méthodes d’harmonisation sur celles-ci. Notre hypothèse était que l’harmonisation devait rapprocher les tendances observées de celles de la littérature. Dans un premier temps, nous avons considéré uniquement les tendances des volumes pour les sujets sains, puis nous avons inclus les sujets atteints de TSA afin d’observer si le statut clinique influence la qualité de l’harmonisation. Dans la majorité des cas, il est simple d’évaluer si l’effet a été positif ou négatif. Par exemple, l’évolution du volume total de matière grise chez les hommes sains est donnée dans [figure 5.2](#). Les données originales sont indiquées en rouge et semblent augmenter avec l’âge. Ceci n’est clairement pas en accord avec les données de référence fondées dans la littérature et indiquées en noir. Les deux méthodes d’harmonisation modifient la tendance et la rapproche de celle de la littérature.

Or, n’ayant pas de métrique quantitative, certaines situations se sont avérées complexes à juger. Comme dans l’exemple [figure 5.2](#), il est difficile de conclure si une méthode est meilleure que l’autre. Au départ, la tendance est inversée par rapport à la littérature. Elle est corrigée après ComBat mais est ‘moins inversée’ après ImUnity. Cependant, il est également possible que les corrections de Combat aient été trop fortes. Afin de pouvoir comparer les méthodes de manière plus précise, nous avons proposé d’utiliser une métrique à partir de l’équation d’une tendance de volume observée et de son équation de référence trouvée dans la littérature. L’idée était d’obtenir une valeur reflétant la similarité des tendances entre les deux équations.

Cette métrique peut être utilisée pour comparer quantitativement les effets d’harmonisation, indépendamment pour chaque ROI. Comme pour la première expérience basée sur l’inspection visuelle, nous avons pu observer des effets d’harmonisation positifs ou négatifs en fonction de la valeur de la métrique. En outre, elle peut également être utilisée pour quantifier le besoin d’harmonisation pour la ROI considérée. C’est un point important car le bruit induit par un site étant probablement non-homogène, certaines zones du cerveau sont plus altérées que d’autres. En outre, cette information peut également être utilisée pour observer tout changement de performance de l’harmonisation par rapport à son besoin initial. Nous nous attendions à avoir un effet positif plus significatif pour les ROIs principalement concernés par l’harmonisation. D’autre part, les tendances des ROIs déjà en phase avec la

littérature sont moins susceptibles d'avoir un tel effet. Il faut toutefois faire attention à ne pas pénaliser les tendances qui sont déjà en accord avec la littérature.

Notre troisième expérience a consisté en une analyse par modèle linéaire mixte. Nous avons cherché à modéliser les évolutions de volumes et d'épaisseurs corticales, en considérant toutes les variables disponibles : âge, sexe, statut, site. En considérant l'affiliation au site comme un facteur de regroupement, nous avons pu estimer son impact sur la métrique observée (volumes ou épaisseurs). Pour ce faire, nous avons utilisé les tests de Kenward-Roger (KR) (Kenward and Roger, 1997), testant la corrélation entre les variables observées et l'origine des images. Ce test statistique a été effectué pour chaque paramètre sous l'hypothèse H_0 : "le paramètre n'a aucune influence sur la métrique observée (volume ou épaisseur)". Pour chaque test, en fonction de sa p-value associée, nous rejetons ($p < 0.05$) ou non H_0 ($p \geq 0.05$). Notre hypothèse était que l'impact des groupes (affiliation au site) avait un impact significatif sur le développement du cerveau avant l'harmonisation et que cet effet devait disparaître après harmonisation. D'autre part, nous nous attendions à une augmentation de variables biologiques significativement corrélées à ces métriques observées, sachant d'après la littérature, que l'âge, le sexe ou les pathologies comme le TSA ont un fort impact sur le développement cérébral.

Notre dernière expérience a consisté en une 'analyse de groupe' via FreeSurfer sur les sujets sains de ABIDE, en utilisant l'origine du site comme groupe et en intégrant des covariables biologiques (âge et sexe) dans notre modèle. Une 'cluster-wise correction' (Hagler et al., 2006) a été utilisée avec un seuil de 10^{-3} comme recommandé dans Greve and Fischl (2018) pour éviter les clusters faux positifs. Cette procédure permet de visualiser les impacts des groupes (ici l'affiliation au site) sur le développement du cerveau, et de visualiser les zones du cerveau les plus affectées par ces variations et d'observer les effets sur ces clusters. Ceci venait en complément de l'expérience précédente basée sur la métrique dérivée utilisée pour quantifier le besoin d'harmonisation dans différents ROIs.

Les résultats de volumétrie cérébrale ont démontré la capacité d'ImUnity à éliminer les biais liés au site ou au scanner tout en ayant un effet positif sur l'étude des trajectoires de développement cortical et sous-cortical. La première expérience a montré que cela était suffisant pour améliorer les tendances apparentes de l'évolution du volume cérébral et les rapprocher des résultats des grandes études mono-centriques. Ceci a été confirmé par la métrique (équation 5.1), avec une amélioration globale pour la tendance des volumes. La plupart des volumes des ROIs semblaient être affectés par le problème d'harmonisation, et nous avons pu améliorer environ 95% (resp. 88%) de l'évolution du volume des ROIs chez les sujets sains (resp. les hommes atteints de TSA). Ces résultats ont été confirmés par l'approche du modèle linéaire mixte. Cela a permis de mettre en évidence la nécessité d'une harmonisation. Le site avait un impact significatif sur l'évolution du volume de la plupart des ROI avant harmonisation. Celui-ci a été réduit après l'harmonisation. D'autre part, l'harmonisation semble avoir réduit les corrélations biologiques, ce qui va à l'encontre de notre hypothèse initiale. ImUnity et ComBAT ont montré des résultats similaires mais ComBat a supprimé plus de corrélations biologiques et a corrigé moins de tendances de volume des ROIs que ImUnity.

Les résultats concernant l'épaisseur corticale ont été moins encourageants. Les deux méthodes n'ont pas significativement supprimer les effets site et semblent avoir du mal à préserver les corrélations biologiques. De plus, l'harmonisation semble avoir dégradé les tendances de certaines ROIs. Nos résultats ne sont pas en accord avec les études précédentes sur le

sujet : (Beer et al., 2020; Fortin et al., 2018). Cependant, il est important de noter la manière différente d'utiliser ComBat dans notre étude où nous avons harmonisé d'abord l'intensité des images puis extrait les métriques d'intérêt (ici les volumes et épaisseurs du cerveau) alors que les métriques sont utilisées directement dans ComBat dans les études précédentes. De plus, nous avons pu observer un besoin d'harmonisation moins marqué concernant les épaisseurs corticales, ce qui peut expliquer que nous n'observons pas d'effet significativement positif de l'harmonisation sur ces observations.

7.6 Conclusion et perspectives

Plusieurs limitations de notre travail peuvent être notées. Premièrement, la solution que nous avons proposée est une configuration 2,5D. Bien qu'elle nécessite moins de paramètres à ajuster qu'une solution 3D, elle tend également à produire des résultats de moindre qualité. Deuxièmement, ImUnity n'a été testé que pour harmoniser des images anatomiques T1w et nous n'avons pas testé le modèle sur d'autres types de séquences. Théoriquement, notre méthode devrait produire des résultats similaires (puisque'elle n'a pas été conçue spécifiquement pour les données T1w), mais il pourrait être très intéressant d'évaluer ses performances sur des images de diffusion IRM ou des images produites par des scanners X. À terme, ImUnity devrait également permettre l'harmonisation multi-séquences (par exemple, T1w et T2w). Il s'agit d'un véritable défi qui n'a pas encore été étudié dans la littérature mais qui pourrait avoir un impact majeur sur les pratiques cliniques. Nous avons commencé une étude sur une harmonisation combinée de T1w et de tomodensitométrie X (CT), et les résultats préliminaires prometteurs justifient des investigations supplémentaires.

De plus, les distorsions géométriques devraient être incluses en plus des variations de contraste. Les derniers résultats présentés (section 5.3) suggèrent en effet que toutes les ROIs ne sont pas impactées de la même manière et que l'alignement des contrastes n'était pas suffisant pour supprimer la totalité du signal induit par les sites. Ceci suggère que l'hypothèse globale selon laquelle ces variations n'impactent que le contraste des images et non leur géométrie est remise en cause.

En plus des développements techniques, des recherches plus approfondies sur les trajectoires de développement du cerveau doivent être effectuées. Si les résultats sont clairs pour l'évolution du volume des ROIs du cerveau, les résultats pour l'épaisseur corticale sont insuffisants pour conclure. Nous notons que la distribution d'âge (figure 5.1) des sujets n'était pas uniforme, avec une distribution Gaussienne centrée autour de 15 ans. L'ajout de sujets provenant d'autres bases de données pourrait uniformiser la distribution pour chaque population et réduire l'intervalle de confiance des régressions. Un autre point critique mentionné dans section 5.4, est l'étude de l'impact biologique sur chaque RCI indépendant. En se basant sur les résultats de la littérature provenant de grands résultats mono-centriques, nous devrions être en mesure de mieux estimer l'impact de chaque caractéristique biologique sur les volumes ou épaisseurs observés et de compter réellement le nombre de ROIs impactées positivement ou négativement par les méthodes d'harmonisation.

En conclusion, il semble que les outils développés au cours de notre projet peuvent être utilisés pour analyser les données DEFIDIAG lorsqu'elles seront disponibles. Même si de nouvelles fonctionnalités peuvent être ajoutées et que des validations supplémentaires peuvent être effectuées, ImUnity semble être en mesure de fournir des images de haute qualité dans

toutes les directions, de supprimer les biais liés aux sites et aux scanners, et d'améliorer la classification des patients et les trajectoires des volumes de développement cérébral dans les grandes et petites régions du cerveau. Nos travaux ont été soumis pour publication dans un journal et présentés dans plusieurs conférences nationales et internationales. Par ailleurs, des collaborations avec des instituts français (CREATIS à Lyon et ARAMIS à Paris) et des laboratoires américains ont été établies pour partager et tester davantage les potentiels de notre approche.

Chapter 8

Personal information

8.1 My resume

Stenzel Cackowski

Grenoble Institute of Neurosciences, La Tronche, 38700

☎ +33 07-66-69-61-04 | ✉ stenzel.cackowski@univ-grenoble-alpes.fr | 📺 Stenzel Cackowski | 🐦 @Cackowski_S

Education

Grenoble Institut of Neurosciences - MIAI

Grenoble

PHD STUDENT - AI AND MRI ANALYSIS TOOLS TO CHARACTERIZE BRAIN ALTERATIONS:

2019 - present

APPLICATIONS TO INTELLECTUAL DISABILITY DISORDERS - MEDICAL IMAGING HARMONIZATION

- Advisor: Dr. Emmanuel L. Barbier
- Advisor: Dr. Michel Dojat
- Advisor: Dr. Thomas Christen

Grenoble INP - ENSIMAG - University Grenoble Alpes

Grenoble

MASTER 2 OF SCIENCE IN INDUSTRIAL AND APPLIED MATHEMATICS - DATA SCIENCES

2016 - 2019

- Honors

Bordeaux INP

Bordeaux

SCIENTIFIC PREPARATORY CLASS

2014 - 2016

- Special subjects : Mathematics, IT

Professional Experience

- | | |
|--------------------------|--|
| June 2022 | Poster presentation about MRI harmonization during OHBM, international conference, Glasgow |
| May 2022 | Oral presentation about MRI harmonization during ISMRM, international conference, London |
| November 2021 | Oral presentation about MRI harmonization during 3IA doctoral workshop, Toulouse |
| September & October 2021 | Oral presentation about MRI harmonization during two conferences : SFRMBM & Workshop 3IA, Grenoble |
| October 2020 | My thesis in 180 seconds - MIAI DAY - Festival Transfo, Grenoble |
| October 2019 | 1st place of "JFR 2019 Data Challenge" - Predicting EDSS score from FLAIR MRI, Pixyl, Grenoble |
| March - August 2019 | 6 months internship - Development of deep-learning models for MRI organs segmentation and CT scans generation, CSIRO Brisbane, Australia |
| Summer 2018 | Engineer Assistant Internship - Image classification project using Deep-Learning in a High Performance Computing environment, Atos, Grenoble |
| 2018 | Research project - Eclipse Attack on Ethereum - Report published on Ensimag's "Ensiwiki", University Grenoble Alpes |
| Summer 2017 | Head of backend development for a Start-up's website, Nsigma, Grenoble |

8.2 Scientific production

Journal papers

Roca P., Attye A., Colas L., Rubini P., Cackowski S. 2020, *Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI*. **Diagnostic and Interventional Imaging**, [10.1016/j.diii.2020.05.009](https://doi.org/10.1016/j.diii.2020.05.009).

Yauy K. et al., 2022, *Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene-phenotype reassessment in clinical routine*. **Genetics in Medicine**, [10.1016/j.gim.2022.02.008](https://doi.org/10.1016/j.gim.2022.02.008).

Submitted journal paper

Cackowski S., Dojat M., Barbier E., Christen T. 2021, *ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization*, in revision, **Medical Imaging Analysis**

Conference abstract

Cackowski S., Dojat M., Barbier E., Christen T. 2022, *A generic solution for multicenter MRI datasets harmonization*, **OHBM 2022**

Conference talks

Cackowski S., Dojat M., Barbier E., Christen T. 2022, *ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization*, **ISMRM 2022**, New Deep Learning Techniques session

Cackowski S., Dojat M., Barbier E., Christen T. 2021, *ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization*, **SFRMBM 2021**, **Workshop 3IA** in Grenoble, **3IA doctoral workshop** in Toulouse

Cackowski S., Dojat M., Barbier E., Christen T. 2020, *ComBat versus CycleGAN for multi-center MR images harmonization*, **SFRMBM 2020**

Awards

2020 My thesis in 180 seconds - Winner MIAI - MIAI-DAY

2019 JFR 2019 Data challenge - Winner Société Française de Radiologie, Pixyl

Teaching

January - July, 2022	Supervision of a 6 months internship - Constance Sohler	Supervisor	<i>Grenoble Institut of Neuroscience - Phelma</i>
10-14th January, 2022	Deep learning for brain MRI: Adversarial generation, constrastive and transfer learning	Teacher	<i>AI4Health school - Practical session</i>
2020-21	Software engineering project - Compiler development - 2*50h	Teacher	<i>ENSIMAG</i>
2021	C Programming course - 28h	Teacher	<i>Phelma</i>

Chapter 9

Supplementary Material

CONTENTS

9.1	JFR-2019 Data challenge with Pixyl	118
9.2	DI-Generic: A deep-learning framework dedicated to medical imaging analyse	127
9.3	Colaboration with Kévin Yauy	129
9.4	Teaching & supervising experience	142

9.1 JFR-2019 Data challenge with Pixyl

At the beginning of my PhD project, I joined the [Pixyl](#) team in the JFR2019 challenge. Pixyl is a French startup from Grenoble developing AI patient care solutions for different pathologies like Multiple Sclerosis (MS). The challenge was held in Paris between September and October 2019, and consisted in predicting MS patient's 'Expanded Disability Status Scale' (EDSS) score based on a FLAIR acquisition as well as several clinical features such as age or sex. EDSS is a score quantifying disability in multiple sclerosis, starting from 0 (so MS symptom) to 10 (death due to MS). According to [wikipedia](#), 'EDSS steps 1.0 to 4.5 refer to people with MS who are fully ambulatory while steps 5.0 to 9.5 are defined by the impairment to ambulation'. The challenge can be seen as a regression problem, trying to predict an increasing non linear metric. The challenge organizers provided us with data to use to train our solution. Then, a different dataset (testing) was given at the time of the challenge and each team had 3 hours to turn in their predictions.

As a computer science engineer, my work was to develop a deep learning learning model taking the FLAIR image, the mask of MS lesions previously segmented by Pixyl solutions and the age of the patient considered as input. I chose to develop a classical Convolutional Neural Network (CNN) that learned features extractions from the FLAIR image and the associated lesions mask. Then the extracted features were concatenated with the biological ones and fed to a tiny dense neural network that made the final prediction.

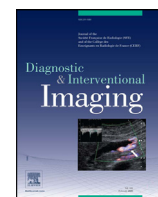
The model's prediction was then pooled with other models (random forest, SVM) results that had been developed by other members of the team. By gathering all these predictions, we trained a last model to make the final prediction. This procedure is known as 'Ensemble learning'. This was motivated by the fact that each model appeared to have complementary performance with the others.

On top of developing efficient solutions, we had to be able to process all the testing data within the 3 hours imposed by the organizers. We thus deployed our trained solutions on several virtual machines that we used in parallel to compute the predictions.

Eventually, we came first in the competition, in front of competing teams like Nvidia or Icometrix. Our solution made the best EDSS predictions in terms of MAE. Moreover, our team was the first to process all the data, which took about 1 hour. This last point is quite impressive when competing with Nvidia's teams, for example, which had massive computing power at their disposal.

As a result of this first place, a paper detailing this experience was published ([Roca et al., 2020](#)). It is presented below.

Overall, this was a very rewarding experience which gave me the opportunity to meet research engineers working in a startup with similar interests to mine. I could feel the pressure for the startup members during the development phase of our solution. In fact, as they mobilized most of the employees for that challenge, they expected the first place or nothing. It was a big bet for a small startup like Pixyl at that time. Ever since, I have kept in touch with Pixyl members who have always treated me very well.



Original article/Computer developments

Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI

P. Roca^{a,*}, A. Attys^{b,c}, L. Colas^d, A. Tucholka^a, P. Rubini^a, S. Cackowski^e, J. Ding^d, J.-F. Budzik^d, F. Renard^{f,g}, S. Doyle^a, E.L. Barbier^e, I. Bousaid^h, R. Casey^{i,j,k,l}, S. Vukusic^{i,j,k,l}, N. Lassau^{m,n}, S. Verclitte^d, F. Cotton^{k,o,p}, On behalf of OFSEP Investigators:¹

^a Pixyl, Research and Development Laboratory, 38000 Grenoble, France

^b Grenoble Alpes University, 38000 Grenoble, France

^c Sydney Imaging Lab, Sydney University, 2006 Sydney, NSW, Australia

^d Imaging Department, Lille Catholic Hospitals, Lille Catholic University, 59000 Lille, France

^e University Grenoble Alpes, Inserm, U1216, Grenoble Institute Neurosciences, 38000 Grenoble, France

^f University Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

^g University Grenoble Alpes, AGEIS, 38000 Grenoble, France

^h Direction Transformation Numérique et Systèmes d'Information, Institut Gustave Roussy, 94805 Villejuif, France

ⁱ Department of Neurology–Multiple Sclerosis, Pathologies de la myéline et neuro-inflammation, Hôpital Pierre Wertheimer, Hospices Civils de Lyon, 69500 Bron, France

^j Université Claude Bernard Lyon 1, Université de Lyon, 69622 Villeurbanne, France

^k Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, 69003 Lyon, France

^l Eugène Devic EDMUS Foundation Against Multiple Sclerosis, 69500 Bron, France

^m Radiology Department, Institut Gustave Roussy, 94805 Villejuif, France

ⁿ BIOMAPS, UMR1281, Université Paris-Saclay, Inserm, CNRS, CEA, Laboratoire d'Imagerie Biomédicale Multimodale Paris-Saclay, 94800 Villejuif, France

^o Department of Radiology, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, 69310 Pierre-Bénite, France

^p CREATIS, CNRS UMR 5220, INSERM U1044, 69622 Villeurbanne, France

ARTICLE INFO

Keywords:

Artificial intelligence
Machine learning
Multiple sclerosis
Disability prediction
Magnetic resonance imaging (MRI)

ABSTRACT

Purpose: The purpose of this study was to create an algorithm that combines multiple machine-learning techniques to predict the expanded disability status scale (EDSS) score of patients with multiple sclerosis at two years solely based on age, sex and fluid attenuated inversion recovery (FLAIR) MRI data.

Materials and methods: Our algorithm combined several complementary predictors: a pure deep learning predictor based on a convolutional neural network (CNN) that learns from the images, as well as classical machine-learning predictors based on random forest regressors and manifold learning trained using the location of lesion load with respect to white matter tracts. The aggregation of the predictors was done through a weighted average taking into account prediction errors for different EDSS ranges. The training dataset consisted of 971 multiple sclerosis patients from the “Observatoire français de la sclérose en plaques” (OFSEP) cohort with initial FLAIR MRI and corresponding EDSS score at two years. A test dataset (475 subjects) was provided without an EDSS score. Ten percent of the training dataset was used for validation.

Abbreviations: 2D, Two-dimensional; 3D, Three-dimensional; AI, Artificial intelligence; CNN, Convolutional neural network; EDSS, Expanded disability status scale; FLAIR, Fluid attenuated inversion recovery; MNI, Montreal Neurological Institute; MRI, Magnetic resonance imaging; MS, Multiple sclerosis; MSE, Mean square error; OFSEP, Observatoire français de la sclérose en plaques; UMAP, Uniform manifold approximation and projection; WMH, White matter hyperintensities.

* Corresponding author at: Pixyl, Research and Development Laboratory, 38000 Grenoble, France.

E-mail address: contact@pixyl.ai (P. Roca).

¹ On behalf of OFSEP Investigators: Steering Committee B. Brochet (Centre hospitalier universitaire de Bordeaux, Hôpital Pellegrin, Service de neurologie, Bordeaux, France), R. Casey (Observatoire français de la sclérose en plaques (OFSEP), Centre de coordination national, Lyon/Bron, France), F. Cotton (Hospices civils de Lyon, Hôpital Lyon sud, Service d'imagerie médicale et interventionnelle, Lyon/Pierre-Bénite, France), J. De Sèze (Hôpitaux universitaires de Strasbourg, Hôpital de Hautepierre, Service des maladies inflammatoires du système nerveux – neurologie, Strasbourg, France), P. Douek (Union pour la lutte contre la sclérose en plaques (UNISEP), Ivry-sur-Seine, France), F. Guillemin (CIC 1433 Epidémiologie Clinique, Centre hospitalier régional universitaire de Nancy, Inserm et Université de Lorraine, Nancy, France), D. Laplaud (Centre hospitalier universitaire de Nantes, Hôpital nord Laennec, Service de neurologie, Nantes/Saint-Herblain, France), C. Lebrun-Frenay (Centre hospitalier universitaire de Nice, Université Nice Côte d'Azur, Hôpital Pasteur, Service de neurologie, Nice, France), L. Mansuy (Hospices civils de Lyon, Département de la recherche clinique et de l'innovation, Lyon, France), T. Moreau (Centre hospitalier universitaire Dijon Bourgogne, Hôpital François Mitterrand, Service de neurologie, maladies inflammatoires du système nerveux et neurologie générale, Dijon, France), J. Olaiz (Université Claude Bernard Lyon 1, Lyon ingénierie projets, Lyon, France), J. Pelletier (Assistance publique des hôpitaux de Marseille, Centre hospitalier de la Timone, Service de neurologie et unité neuro-vasculaire, Marseille, France), C. Rigaud-Bully (Fondation Eugène Devic EDMUS contre la sclérose en plaques, Lyon, France), B. Stankoff (Assistance publique des hôpitaux de Paris, Hôpital Saint-Antoine, Service de neurologie, Paris, France),

<https://doi.org/10.1016/j.diii.2020.05.009>

2211-5684/© 2020 Société française de radiologie. Published by Elsevier Masson SAS. All rights reserved.

Results: Our algorithm predicted EDSS score in patients with multiple sclerosis and achieved a MSE = 2.2 with the validation dataset and a MSE = 3 (mean EDSS error = 1.7) with the test dataset.

Conclusion: Our method predicts two-year clinical disability in patients with multiple sclerosis with a mean EDSS score error of 1.7, using FLAIR sequence and basic patient demographics. This supports the use of our model to predict EDSS score progression. These promising results should be further validated on an external validation cohort.

© 2020 Société française de radiologie. Published by Elsevier Masson SAS. All rights reserved.

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system, which remains the leading cause of non-traumatic disability in young people and is associated with a high economic burden on society partly due to the high cost of the available treatments [1,2]. In most patients with MS, the initial phase of the disease consists of reversible episodes of neurological deficits and over time, the development of permanent neurological deficits and progression of clinical disability [3]. Correctly predicting short-term outcome in patients with MS is an important issue as this could help identify patients who may benefit from a more aggressive treatment.

So far, knowledge of natural disability evolution of MS is mainly based on cohort studies and focused on long-term clinical progression. As a consequence, baseline factors strongly predictive of worsening disability have not yet been fully identified [4,5]. Advanced statistical modeling using support vector machine and

random forest has been recently applied on 1582 patients to predict short-term expanded disability status scale (EDSS) score progression after 2 years from a comprehensive list of baseline factors [6]. These factors included clinical factors (such as age, gender, ethnicity, number of relapses 1 and 3 years prior to study, disease duration, prior treatment, EDSS score) and imaging factors (number of lesions, lesion volume and brain parenchymal fraction) [6]. Nevertheless, the predictor models showed poor discriminating capabilities so that there is a need for alternate predictors [6].

Artificial intelligence (AI) has demonstrated utility in the identification of abnormalities on imaging studies [7–11]. However, the capabilities of AI as directly applied to fluid-attenuated inversion recovery (FLAIR) MRI data have received little attention so far because of the well-known “clinico-radiological paradox”. Indeed, the clinical course of MS based on the burden of lesions is known to be unpredictable [12–14]. It is not clear whether this paradox relies on a lack of information, for example regarding the gray matter MS

S. Vukusic (Hospices civils de Lyon, Hôpital Pierre Wertheimer, Service de neurologie A, Lyon/Bron, France), , Investigators R. Marignier (Hospices civils de Lyon, Hôpital Pierre Wertheimer, Service de neurologie A, Lyon/Bron, France), M. Debouverie (Centre hospitalier régional universitaire de Nancy, Hôpital central, Service de neurologie, Nancy, France), G. Edan (Centre hospitalier universitaire de Rennes, Hôpital Pontchaillou, Service de neurologie, Rennes, France), J. Ciron (Centre hospitalier universitaire de Toulouse, Hôpital Purpan, Service de neurologie inflammatoire et neuro-oncologie, Toulouse, France), A. Ruet (Centre hospitalier universitaire de Bordeaux, Hôpital Pellegrin, Service de neurologie, Bordeaux, France), N. Collongues (Hôpitaux universitaires de Strasbourg, Hôpital de Hautepierre, Service des maladies inflammatoires du système nerveux – neurologie, Strasbourg, France), C. Lubetzki (Assistance publique des hôpitaux de Paris, Hôpital de la Pitié-Salpêtrière, Service de neurologie, Paris, France), P. Vermersch (Centre hospitalier universitaire de Lille, Hôpital Salengro, Service de neurologie D, Lille, France), P. Labauge (Centre hospitalier universitaire de Montpellier, Hôpital Gui de Chauliac, Service de neurologie, Montpellier, France), G. Defer (Centre hospitalier universitaire de Caen Normandie, Service de neurologie, Hôpital Côte de Nacre, Caen, France), M. Cohen (Centre hospitalier universitaire de Nice, Université Nice Côte d’Azur, Hôpital Pasteur, Service de neurologie, Nice, France), A. Fromont (Centre hospitalier universitaire Dijon Bourgogne, Hôpital François Mitterrand, Service de neurologie, maladies inflammatoires du système nerveux et neurologie générale, Dijon, France), S. Wiertlewsky (Centre hospitalier universitaire de Nantes, Hôpital nord Laennec, Service de neurologie, Nantes/Saint-Herblain, France), E. Berger (Centre hospitalier régional universitaire de Besançon, Hôpital Jean Minjot, Service de neurologie, Besançon, France), P. Clavelou (Centre hospitalier universitaire de Clermont-Ferrand, Hôpital Gabriel-Montpied, Service de neurologie, Clermont-Ferrand, France), B. Audoin (Assistance publique des hôpitaux de Marseille, Centre hospitalier de la Timone, Service de neurologie et unité neuro-vasculaire, Marseille, France), C. Giannesini (Assistance publique des hôpitaux de Paris, Hôpital Saint-Antoine, Service de neurologie, Paris, France), O. Gout (Fondation Adolphe de Rothschild de l’œil et du cerveau, Service de neurologie, Paris, France), E. Thouvenot (Centre hospitalier universitaire de Nîmes, Hôpital Carémieu, Service de neurologie, Nîmes, France), O. Heinzlef (Centre hospitalier intercommunal de Poissy Saint-Germain-en-Laye, Service de neurologie, Poissy, France), A. Al-Khedr (Centre hospitalier universitaire d’Amiens Picardie, Site sud, Service de neurologie, Amiens, France), B. Bourre (Centre hospitalier universitaire Rouen Normandie, Hôpital Charles-Nicolle, Service de neurologie, Rouen, France), O. Casez (Centre hospitalier universitaire Grenoble-Alpes, Site nord, Service de neurologie, Grenoble/La Tronche, France), P. Cabre (Centre hospitalier universitaire de Martinique, Hôpital Pierre Zobda-Quitman, Service de Neurologie, Fort-de-France, France), A. Montcuquet (Centre hospitalier universitaire Limoges, Hôpital Dupuytren, Service de neurologie, Limoges, France), A. Créange (Assistance publique des hôpitaux de Paris, Hôpital Henri Mondor, Service de neurologie, Créteil, France), J.-P. Camdessanché (Centre hospitalier universitaire de Saint-Étienne, Hôpital Nord, Service de neurologie, Saint-Étienne, France), J. Faure (Centre hospitalier universitaire de Reims, Hôpital Maison-Blanche, Service de neurologie, Reims, France), A. Maurousset (Centre hospitalier régional universitaire de Tours, Hôpital Bretonneau, Service de neurologie, Tours, France), I. Patry (Centre hospitalier sud francilien, Service de neurologie, Corbeil-Essonnes, France), K. Hankiewicz (Centre hospitalier de Saint-Denis, Hôpital Casanova, Service de neurologie, Saint-Denis, France), C. Pottier (Centre hospitalier de Pontoise, Service de neurologie, Pontoise, France), N. Maubeuge (Centre hospitalier universitaire de Poitiers, Site de la Milétrie, Service de neurologie, Poitiers, France), C. Labeyrie (Assistance publique des hôpitaux de Paris, Hôpital Bicêtre, Service de neurologie, Le Kremlin-Bicêtre, France), C. Nifle (Centre hospitalier de Versailles, Hôpital André-Mignot, Service de neurologie, Le Chesnay, France), , Imaging group R. Ameli (Hospices civils de Lyon, Service de radiologie, Lyon, France), R. Anxionnat (CHU Nancy, Service de radiologie, Nancy, France), A. Attye (CHU de Grenoble, Service de radiologie, Grenoble, France), E. Bannier (Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France), C. Barillot (INRIA, Rennes, France), D. Ben Salem (CHU Brest, Service de radiologie, Brest, France), M.-P. Boncoeur-Martel (CHU Limoges, Service de radiologie, Limoges, France), F. Bonneville (CHU Toulouse Purpan, Service de radiologie, Toulouse, France), C. Boutet (CHU Saint-Etienne, Service de radiologie, Saint-Etienne, France), J.-C. Briset (Median technologies, Valbonne, France), F. Cervenanski (CREATIS, Villeurbanne, France), B. Claise (CHU Clermont-Ferrand, Service de radiologie, Clermont-Ferrand, France), O. Commowick (NRIA, Rennes, France), J.-M. Constans (CHU Amiens-Picardie, Service de radiologie, Amiens, France), P. Dardel (CH Chambéry, Service de radiologie, Chambéry, France), H. Desal (CHU Nantes, Service de radiologie, Nantes, France), Vincent Dousset (CHU Bordeaux, Service de radiologie, Bordeaux, France), F. Durand-Dubief (Hospices civils de Lyon, Service de Neurologie, Lyon, France), J.-C. Ferre (CHU Rennes, Service de radiologie, Rennes, France), E. Gerardin (CHU Rouen, Service de radiologie, Rouen, France), T. Glattard (CREATIS, Villeurbanne, France), S. Grand (CHU de Grenoble, Service de radiologie, Grenoble, France), T. Grenier (CREATIS, Villeurbanne, France), R. Guillemin (CHU Poitiers, Service de radiologie, Poitiers, France), C. Guttman (Harvard Medical School, Boston, USA), A. Krainik (CHU Grenoble Alpes, Service de radiologie, Grenoble, France), S. Kremer (CHU Strasbourg, Service de radiologie, Strasbourg, France), S. Lion (Centre de coordination national de l’OFSEP, Lyon/Bron, France), N. Menjot de Champfleury (CHU Montpellier, Service de radiologie, Montpellier, France), L. Mondot (CHU Nice, Service de radiologie, Nice, France), O. Outteryck (CHRU Lille, Consultations de neurologie D, Lille, France), N. Pyatigorskaya (ICM, Service de radiologie, Paris, France), J.-P. Pruvo (CHRU Lille, Service de radiologie, Lille, France), S. Rabaste (Hospices civils de Lyon, Service de radiologie, Lyon, France), J.-P. Ranjeva (APHM – CHU Marseille Timone, Service de radiologie, Marseille, France), J.-A. Roch (Hôpital privé Jean Mermoz, Service de radiologie, Lyon, France), J.C. Sadik (Fondation A. de Rothschild, Service de radiologie, Paris, France), D. Sappey-Marini (Hospices civils de Lyon, Service de radiologie, Lyon, France), J. Savatovsky (Fondation A. de Rothschild, Service de radiologie, Paris, France), J.-Y. Tanguy (CH Angers, Service de radiologie, Angers, France), A. Tourbah (Hôpital Raymond Poincaré, Service de Neurologie, Garches, France), T. Tourdias (CHU Bordeaux, Service de radiologie, Bordeaux, France),

Table 1
MRI characteristics of the different datasets.

MRI characteristic	Training set	Validation set	Test set
Three-dimensional	557 (557/856; 65%)	67 (67/96; 70%)	410 (410/475; 86%)
Two-dimensional	299 (299/856; 35%)	29 (29/96; 30%)	65 (65/475; 14%)
3T	445 (445/856; 52%)	58 (58/96; 60%)	237 (237/475; 50%)
1.5T	411 (411/856; 48%)	38 (38/96; 40%)	238 (238/475; 50%)
Siemens/Philips/GE/Canon (%)	42/45.1/12.4/0.5	40/49/11/0	44/41/14.6/0.4
Period of MR image acquisition (years)	2008–2017	2009–2017	2015–2019
Dataset count ^a	856	96	475

Siemens: Siemens Healthineers; Philips: Philips Healthcare; GE: General Electric Healthcare; Canon: Canon Medical Systems.

^a 19 subjects were excluded from training and validation sets due to poor image quality or small field of view.

injuries, or due to the absence of appropriate tools to analyze the white matter spatial distribution of MS lesions.

The purpose of this study was to create an algorithm that combines multiple machine-learning techniques with the ability to predict EDSS score of patients with MS, based on age, sex and FLAIR MRI data.

2. Materials and methods

2.1. Study population

The FLAIR MRI data were provided as part of the “Multiple Sclerosis” challenge organized during the 2019 edition of the Journées françaises de radiologie, which is the annual meeting of the French Society of Radiology (Société française de radiologie). Two training datasets of patients with MS with initial FLAIR MRI and EDSS score at two years were used. A first dataset (DS1) included 480 subjects and a second one (DS2) 491 subjects. A third new test dataset without EDSS values (DS3) of 475 subjects was the reference to evaluate the exactness of the model. Datasets DS1, DS2, and DS3 were part of the OFSEP (“Observatoire français de la sclérose en Plaques”) cohort, registered on clinicaltrials.gov (NCT02889965) and compliant with French data confidentiality regulations.

The MRI characteristics of the different datasets are summarized in Table 1. This multi-centric dataset originated from 37 institutions in 13 French cities and contained a variety of FLAIR

sequences (i.e., two-dimensional [2D] or three-dimensional [3D] acquisition, sagittal/axial or coronal planes, contrast-enhanced or not, and various imaging parameters) acquired using various MRI units (Siemens Healthineers, General Electric Healthcare, Philips Healthcare, Canon Medical Systems) and magnetic fields (1.5 T or 3 T).

2.2. MS disability prediction

Different complementary strategies were combined. They included intensity bias field correction, FLAIR normalization to a customized brain template, data augmentation, tract-based lesion load computation, pre-training, ensemble aggregation of a pure deep learning model [15] and predictor models using “hand-crafted features” based on a priori anatomical knowledge, and parallel deployment on the Pixyl Cloud Infrastructure. Fig. 1 presents the flowchart of our pipeline.

2.3. Preprocessing and training/validation split

FLAIR images were first corrected for inhomogeneities using the N4 algorithm [16] and registered to a common home-made FLAIR template provided by the Montreal Neurological Institute (MNI) using linear and nonlinear registrations of the ANTS library [17]. After this step, FLAIR images were normalized to zero mean and unit variance and resized to $148 \times 148 \times 154$ voxels. The FLAIR

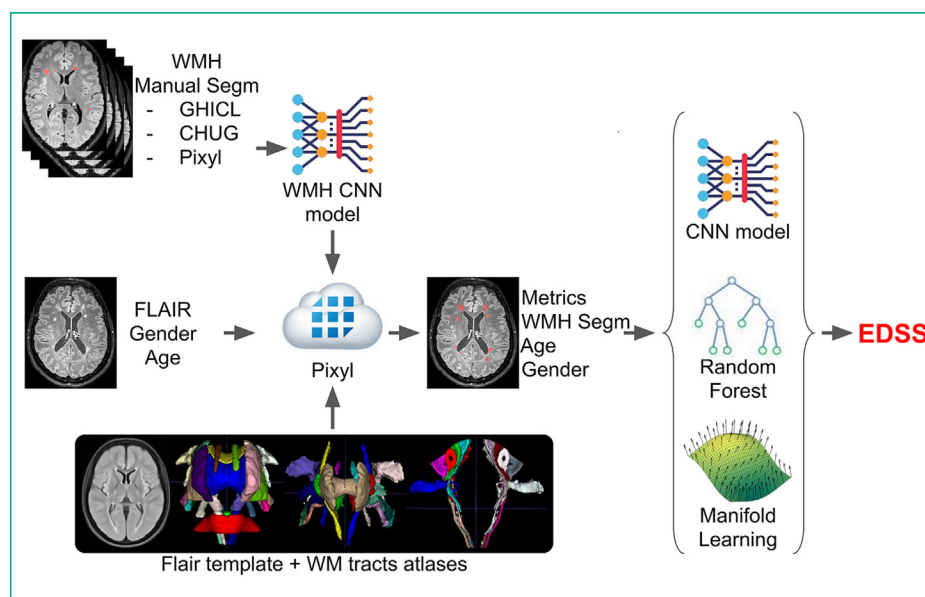


Fig. 1. Flowchart of the prediction pipeline. CNN: convolutional neural network; EDSS: expanded disability status scale; FLAIR: fluid attenuated inversion recovery; GHICL: Groupement des Hôpitaux de l'Institut Catholique de Lille; CHUG: Centre Hospitalier Universitaire de Grenoble.

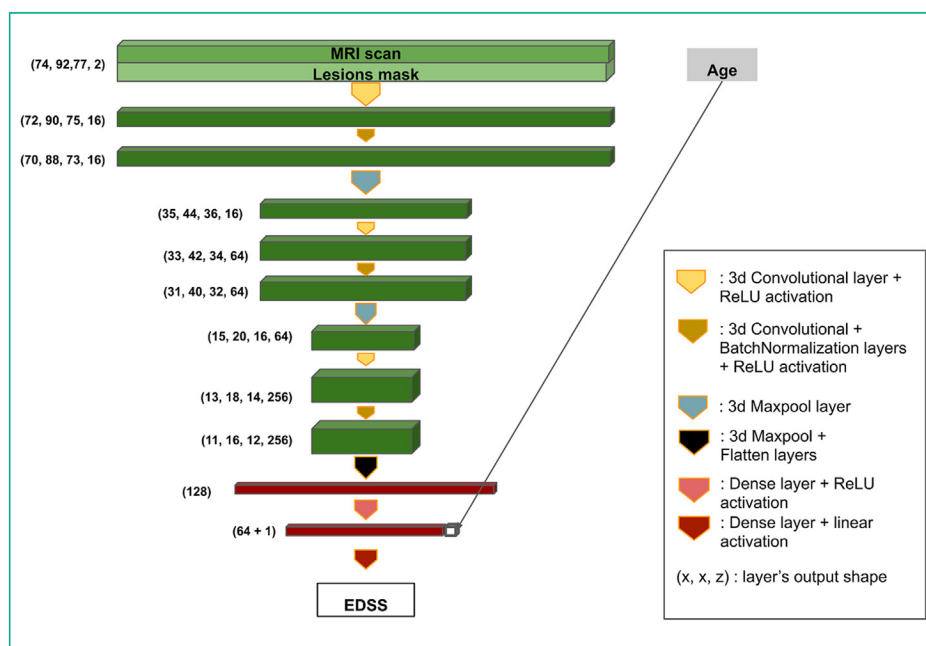


Fig. 2. Drawing shows the architecture of the convolutional neural network predictor. EDSS=expanded disability status scale;.

template was built using a subset of 195 3D FLAIR images from DS1 first registered to the MNI space using an affine transform by two observers (A.T., P. Ro.) [18].

In order to train and validate the predictor models, we divided the union of DS1 and DS2 into a training set (90%) and a validation set (10%) in a stratified way guaranteeing that each subgroup follows the same EDSS distribution.

2.4. Deep learning predictor

To facilitate the prediction task, we divided the problem into two steps. First, we segmented white matter hyperintensities (WMH) from linearly normalized FLAIR MRI, leading to a lesion map in normalized space. Second, we predicted the two-year EDSS score from age, normalized FLAIR MRI and lesion map.

WMH were segmented using the Pixyl.Neuro CE-marked solution (<https://pixyl.ai/>). This solution used a convolutional neural network (CNN) based on a multi-level patch-based series of convolutions and max pools in TensorFlow. The CNN was pre-trained on hundreds of FLAIR images from multiple MRI manufacturers, labeled by experts, augmented using noise, inhomogeneities and geometric deformations. The CNN was retrained using an additional dataset of 29 FLAIR images of MS patients from the “Groupement des Hôpitaux de l’Institut Catholique de Lille” manually segmented by three expert radiologists (S.V., J. D., L. C.).

To predict the EDSS score, a 3D-CNN composed of three convolutional blocks, each corresponding to a succession of two 3D convolutional layers followed by 3D max-pooling layer was developed (Fig. 2). A ReLU activation was added after each convolutional layer and batch-normalization was used after the second convolutional layer. After extraction of the features we added a succession of dense layers. Patient age was added as a new feature in the last layer as it is one of the most relevant features for EDSS score prediction and it would help increase algorithm performance. Finally, this densely connected layer of 65 features predicted the EDSS score. We addressed the EDSS score prediction from a regression point of view as the EDSS scores are ordered by disability severity. Weights were initialized using a truncated normal distribution centered

on 0 with a standard deviation of 0.02. The model was trained on batches of size 16, using Adam optimizer with a learning rate of $10e-3$ to minimize the mean squared error (MSE) loss function.

Two instances of the model were trained: one using the FLAIR and lesion map linearly normalized to the MNI space which takes into account brain atrophy specific to each subject, and a second one based on the non-linear registration computed previously less sensitive to atrophy. Indeed, the non-linear registration allows a better matching of anatomical structures, but can mask relevant differences, particularly those associated with brain atrophy.

2.5. Classical machine learning predictors using anatomical knowledge

A dimension reduction was performed using handcrafted features summarizing the impact of lesions on the brain network through tract-based lesion load computation, and more general volumetric measures. Quantitative analysis of white matter lesion burden was performed in 60 tracts of interest from the ICBM-DTI-81 white matter labels [19–21] and the sensorimotor tracts atlases [22] in MNI space using nonlinear registration. In addition, measures of whole-brain lesion load and volume of the lateral ventricles were performed. These volumetric measures, combined with age, gender and 3D/2D nature of FLAIR sequence constituted 65 features used to train two additional EDSS score predictors.

The first predictor used random forest regressors from scikit-learn [23,24] with 200 estimators and three samples minimum per leaf to reduce overfitting. The random forest regressor was trained twice, firstly on the whole training dataset (RF single) then on a subset containing 3D FLAIR images only. These two models were combined in a unique predictor (“RF dual”) using the 3D nature of the input data (Fig. 3).

A second complementary predictor using manifold learning was built using the uniform manifold approximation and projection (UMAP) algorithm, chosen for its good property to preserve the global structure of the data [25]. Then the EDSS score was predicted in a 2D reduced space by a local interpolation of the targets associated with the nearest neighbors in the training set.

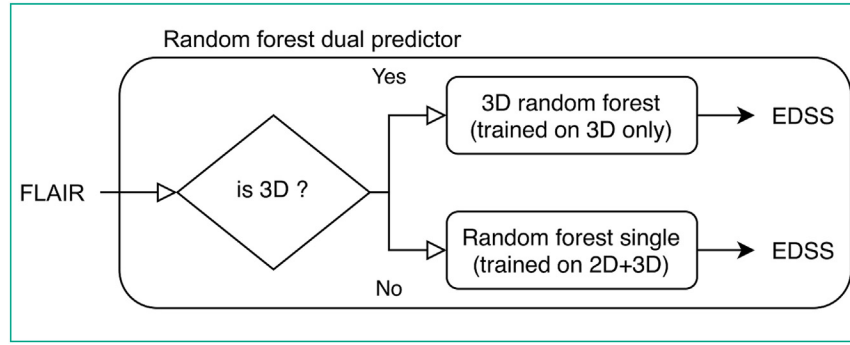


Fig. 3. Flowchart associated with the random forest dual predictor. Depending on the three-dimensional (3D) nature of the input data, the model uses either a model trained exclusively on 3D FLAIR or one trained on both 3D and 2D FLAIR to predict the EDSS score. FLAIR: fluid attenuated inversion recovery; EDSS: expanded disability status scale.

2.6. Ensemble aggregation and implementation

To evaluate the performance of each predictor, we classified EDSS scores in ten groups according to the EDSS integer part: group $p = 0 : EDSS < 1$, group $p = 1 : 1 \leq EDSS < 2$, ..., group $p = 9 : 9 \leq EDSS < 10$, group $p = 10 : EDSS = 10$. For each predictor k the mean square error across each EDSS group ($MSE_k(p)$, for $p = 1 \dots 10$) was computed on the validation dataset. The EDSS score predictors were then aggregated using a weighted average where the weights associated with each predictor relied on its performance on the validation dataset as follows:

For each subject x :

$$edss_{agg}(x) = \sum_k w_k(\text{floor}(edss_k(x))) edss_k(x) \times \left(\frac{1}{\sum_k w_k(\text{floor}(edss_k(x)))} \right)$$

where $edss_k$ is the EDSS score predicted by the k th predictor, and $w_k(p)$, the weight associated with this predictor for the EDSS group p , is equal to the inverse of $MSE_k(p)$ presented previously. In order to study the contribution of this aggregated predictor compared to age only, an additional predictor using Ridge linear regression based on age was built.

To provide the prediction in DS3 within two hours, we integrated all processing steps, from preprocessing to ensemble

aggregation, in an automated pipeline that predicted the EDSS score from a raw FLAIR sequence, as well as age and gender. By having a stand-alone pipeline, we were able to use Pixyl's infrastructure to run all the analyses in parallel.

3. Results

3.1. Datasets characteristics

In addition to the heterogeneity in MRI quality, the EDSS score distribution was very unbalanced (Fig. 4). There were more low scores than high scores (81% of EDSS scores ≤ 4) and > 22% of the samples corresponded to an EDSS score of 0. In addition, integer scores were over-represented (75% of EDSS > 0) by comparison with non-integer scores (25%).

3.2. Score, ranking and predictor performances

We achieved a MSE score of 3 (associated with an estimated mean EDSS score error of 1.7) on this new dataset of 475 subjects and submitted the results in one hour and a half, scoring first in the data challenge. Fig. 5 presents the most informative features, with the associated measure of importance given by the random forest model trained on 3D FLAIR.

Our different predictors demonstrated similar performance in terms of global MSE on the validation dataset (Table 2), with the random forest model specific to 3D/2D data ranking first with a

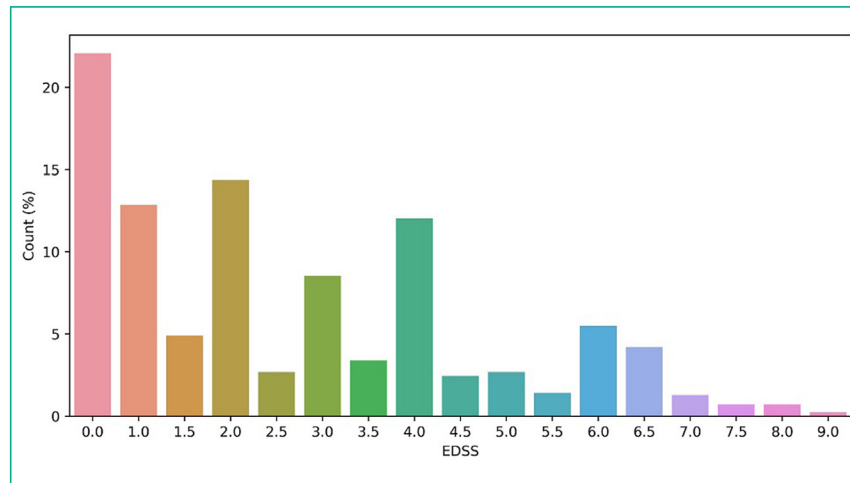


Fig. 4. Histogram of expanded disability status scale scores in the training set, reflecting the unbalanced nature of EDSS distribution. There are more low scores than high scores (81% of EDSS scores inferior or equal to 4) and more than 22% of the samples correspond to an EDSS score of 0. Integer (1, 2, 3, etc.) scores are more represented (75% of EDSS > 0) than non-integer (1.5, 2.5, etc.) scores (25%), this could be due to a human bias towards integer scores when scoring. EDSS: expanded disability status scale.

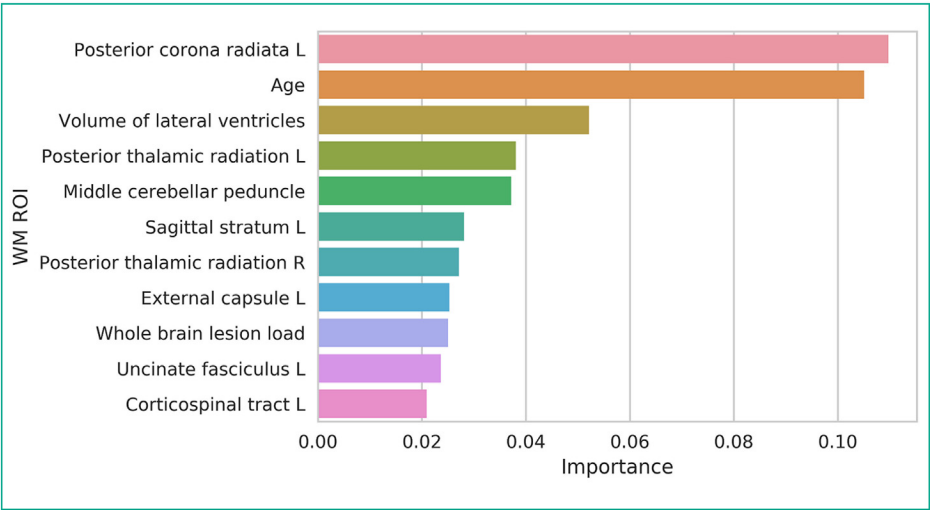


Fig. 5. Diagram shows the most informative features associated with the random forest single predictor model. “L” and “R” mean “Left” and “Right” respectively. WM: white matter; ROI: region of interest.

Table 2
Mean square error of each model with the validation set.

Method	MSE on the validation set	MSE on the test set
Age-only ridge regression	3.779	Unknown
CNN with linear registration	2.705	Unknown
CNN with nonlinear registration	2.714	Unknown
Random forest dual	2.560	Unknown
Random forest single	2.697	Unknown
Manifold	3.216	Unknown
Aggregated model	2.210	3

CNN: Convolutional Neural Network; MSE: mean square error.

MSE of 2.56. The aggregated model reached a MSE of 2.21. For comparison purposes, the Ridge regression model based on age only had a MSE of 3.8. Fig. 6 shows the example of two patients with MS for whom the aggregated model correctly predicted the EDSS scores while the Ridge regression model did not. These two patients (A and B) were 45- and 55-year-old, respectively and had close imaging characteristics at baseline but a different EDSS at two years (3 and 6.5 respectively). Our quantitative image analysis revealed differences between the two patients in terms of volume of lateral ventricles (60 mL and 84 mL for A and B respectively) and left posterior corona radiata lesion load (33% and 48% for A and B respectively) (Fig. 6). We obtained the best predic-

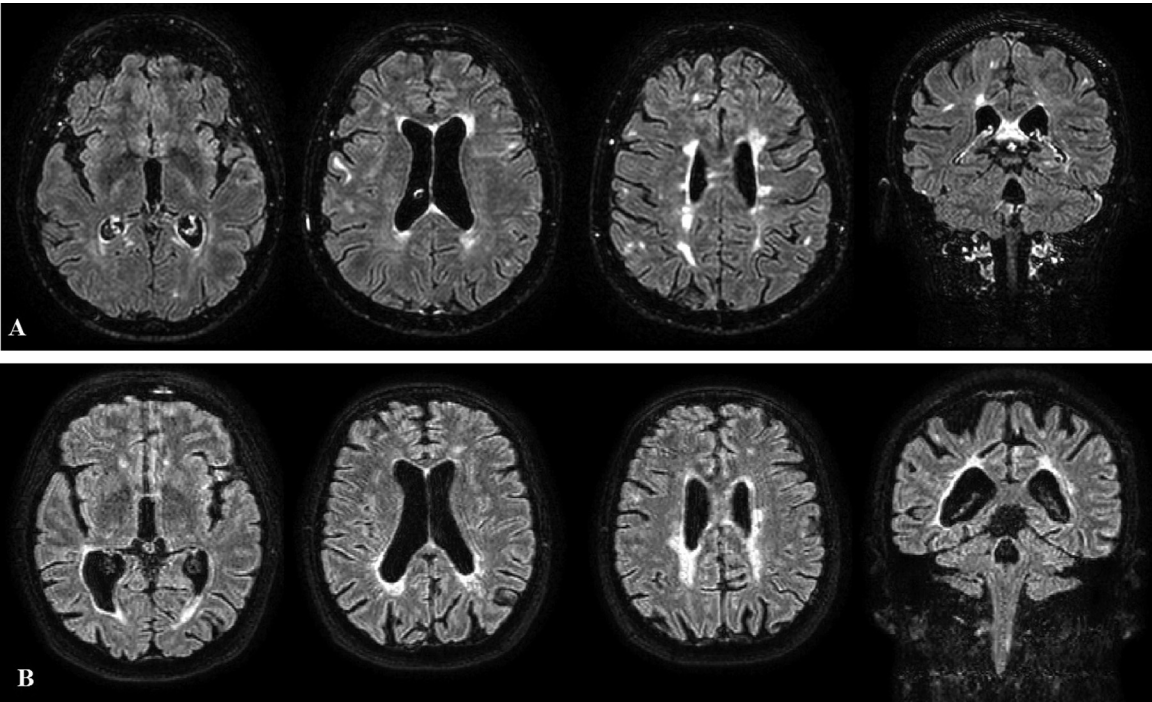


Fig. 6. FLAIR images of two patients with multiple sclerosis (MS) for which the aggregated model correctly predicted the expanded disability status scale scores while the Ridge regression model using age did not. A. 46-year-old woman with MS, volume of lateral ventricles = 60 mL, left posterior corona radiata lesion load = 2.5 mL (33%), EDSS at two years = 3. B. 55-year-old man with MS, volume of lateral ventricles = 84 mL, left posterior corona radiata lesion load = 3.6 (48%), EDSS at two years = 6.5. The age-only Ridge regression model predicted an EDSS of 3 and 3.5 for A and B respectively.

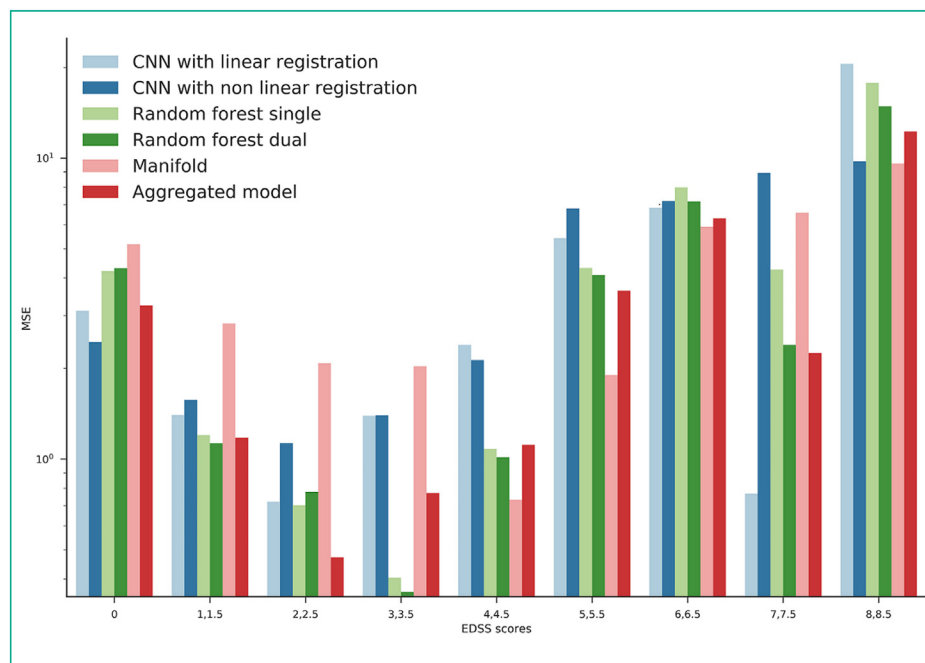


Fig. 7. Graph shows results of each model on the validation set: MSE for each EDSS group. The log-scale was used in order to facilitate the visualization. CNN: convolutional neural network; EDSS: expanded disability status scale; MSE: mean square error.

tion for middle EDSS scores (predictors presenting a MSE < 1.1 when $1 < \text{EDSS score} < 4.5$), and the worst prediction for high EDSS scores (MSE > 9.6 for EDSS score ≥ 8). For EDSS score = 0, no model performed particularly well (MSE superior to 2.4 for all models), despite the relatively large number of training examples ($n = 189$ corresponding to 22% of the training set). Fig. 7 shows the high variability of model performances across EDSS scores of the validation set.

4. Discussion

Our method can predict two-year clinical disability with a mean square EDSS score error of 3 only based on a single, baseline, routine FLAIR MRI examination with some basic clinical information, with heterogeneous imaging quality from various MRI equipments and centers. The most informative variables were the age, the volume of the lateral ventricles, and the lesion load in main white matter tracts such as corona radiata, thalamic radiation, and cerebellar peduncle.

The aggregation of complementary predictor models, through a weighted average taking into account prediction errors for different EDSS score ranges, allowed us to benefit from the strength of each predictor. Indeed, not a single predictor performs well on each one of the EDSS scores. On the contrary, the best predictor varies across EDSS score groups and as expected, the aggregated model presented shows improved performance compared to the best individual predictor. The features characterizing the impact of lesions on the brain network (using tract-based lesion loads) demonstrate their usefulness over features learned using a pure image-based deep learning approach for middle EDSS scores, reaching a very low MSE < 0.36 for EDSS of 3 and 3.5 on the validation set. We also achieve better prediction accuracy (MSE = 2.2) on this dataset compared to an age-only Ridge regression model (MSE = 3.8), highlighting the importance of imaging features in the prediction. Further studies including quantitative metrics coming from T1-weighted-based segmentation could be interesting to understand the influence of atrophy on the clinical disability.

Our study has some limitations. First, our aggregated model had difficulty to predict EDSS score = 0 at two years. The injection

of a priori anatomical knowledge on brain connections was not sufficient to overcome the clinico-radiological paradox for these patients. This could be due to various factors including intra- and inter-variability when scoring EDSS, particularly with low scores [26], underestimation of damage to the normal-appearing brain tissue, neglect of spinal cord involvement, or masking effect of cortical plasticity. Second, for EDSS scores > 8, the aggregated model reached an MSE over 9 on the validation set. This result could be explained by the unbalanced EDSS score distribution of the training set which presented a limited number of examples of these high EDSS scores. A training session on a larger cohort of patients could overcome this limitation or different solutions could be tested to artificially increase the number of high EDSS scores during the training such as oversampling of high EDSS score examples or generating synthetic data. Last, the initial EDSS scores (associated with the baseline MRI examination) were not available during this challenge, thus making it impossible to estimate clinical disability progression over the follow-up. In addition, we received no information about patient treatment, so we were not able to study therapeutic effects on our disability prediction.

In conclusion, our model helps predict the EDSS score at two years for patients affected by MS by relying solely on a single FLAIR sequence and basic demographic information. This performance was achieved through the combination of multiple predictors based on images, anatomical priors, and white matter lesion load using MRI from multiple clinical centers. These promising results should be further validated on an external larger test cohort and have the potential to be highly relevant for disability prediction and the evaluation of disease-modifying treatments. This supports the use of our model to predict EDSS score progression and/or the improvement of the current prediction using additional factors such as baseline EDSS score.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

CRediT authorship contribution statement

Pauline Roca: conceptualization, data curation, formal analysis, methodology, writing- original draft, writing- review & editing; Arnaud Attyé: investigation, writing- original draft, writing- review & editing; Lucie Colas: resources, writing- review & editing; Alan Tucholka: conceptualization, data curation, methodology, supervision, writing- original draft, writing- review & editing; Pascal Rubini: conceptualization, data curation, formal analysis, methodology, software, writing- original draft, writing- review & editing; Stenzel Cackowski: formal analysis, writing- original draft, writing- review & editing; Juliette Ding: resources, writing- review & editing; Jean-François Budzik: resources, writing- review & editing; Felix Renard: formal analysis, writing- original draft, writing- review & editing; Senan Doyle: resources, writing- review & editing; Emmanuel L. Barbier: resources, writing- review & editing; Imad Bousaid: investigation; Romain Casey: investigation, resources, writing- review & editing; Sandra Vukusic: investigation, resources, writing- review & editing; Nathalie Lassau: investigation, resources, writing- review & editing; Sébastien Vercllytte: conceptualization, investigation, resources, writing- original draft, writing- review & editing; and François Cotton: investigation, resources, writing- review & editing.

Acknowledgements

This work was conducted using data from the Observatoire Français de la Sclérose en Plaques (OFSEP) which is supported by a grant provided by the French State and handled by the “Agence Nationale de la Recherche” within the framework of the “Investments for the Future” program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation.”

Disclosure of interest

Pauline Roca, Alan Tucholka, and Pascal Rubini are employees at Pixyl. Arnaud Attyé is a part-time consultant at Pixyl. Felix Renard has a grant from Carnot-LSI for work unrelated to the contents of

this manuscript. Lucie Colas, Stenzel Cackowski, Juliette Ding, Jean-François Budzik, Emmanuel Barbier, Imad Bousaid, Romain Casey, Sandra Vukusic, Nathalie Lassau, Sébastien Vercllytte, and François Cotton declare that they have no competing interest.

References

- [1] Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, et al. Cost of disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol* 2011;21:718–79.
- [2] Chen AY, Chonghasawat AO, Leadholm KL. Multiple sclerosis: frequency, cost, and economic burden in the United States. *J Clin Neurosci* 2017;45:180–6.
- [3] Filippi M. Multiple sclerosis. *Nat Rev Dis Primer* 2018;4:27.
- [4] Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH, et al. Change in MS-related disability in a population-based cohort: a 10-year follow-up study. *Neurology* 2004;62:51–9.
- [5] Jokubaitis VG, Spelman T, Kalincik T, Lorscheider J, Havrdova E, Horakova D, et al. Predictors of long-term disability accrual in relapse-onset multiple sclerosis. *Ann Neurol* 2016;80:89–100.
- [6] Pellegrini F, Copetti M, Sormani MP, Bovis F, de Moor C, Debray TP, et al. Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. *Mult Scler J* 2019 [135245851988734].
- [7] Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagn Interv Imaging* 2019;100:327–36.
- [8] Herent P, Schmauch B, Jehanno P, Dehaene O, Saillard C, Baileysguier C, et al. Detection and characterization of MRI breast lesions using deep learning. *Diagn Interv Imaging* 2019;100:219–25.
- [9] Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagn Interv Imaging* 2019;100:235–42.
- [10] Lassau N, Estienne T, de Vomécourt P, Azoulay M, Cagnol J, Garcia G, et al. Five simultaneous artificial intelligence data challenges on ultrasound, CT, and MRI. *Diagn Interv Imaging* 2019;100:199–209.
- [11] Group SFR-IA, French Radiology Community CERF. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. *Diagn Interv Imaging* 2018;99:727–42.
- [12] Mollison D, Sellar R, Bastin M, Mollison D, Chandran S, Wardlaw J, et al. The clinico-radiological paradox of cognitive function and MRI burden of white matter lesions in people with multiple sclerosis: a systematic review and meta-analysis. *PLoS One* 2017;12:e0177727.
- [13] Altermatt A, Gaetano L, Magon S, et al. Clinical correlations of brain lesion location in multiple sclerosis: voxel-based analysis of a large clinical trial dataset. *Brain Topogr* 2018;31:886–94.
- [14] Barkhof F. MRI in multiple sclerosis: correlation with expanded disability status scale (EDSS). *Mult Scler J* 1999;5:283–6.
- [15] Rokach L. Ensemble-based classifiers. *Artif Intell Rev* 2010;33:1–39.
- [16] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 2010;29:1310–20.
- [17] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 2011;54:2033–44.
- [18] Avants BB, Tustison N, Song G. Advanced normalization tools (ANTs). *Insight J* 2009;2:1–35.
- [19] Mori S, Wakana S, Van Zijl PC, Nagae-Poetscher LM. MRI atlas of human white matter. Elsevier; 2005.
- [20] Wakana S, Caprihan A, Panzenboeck MM, Fallon JH, Perry M, Gollub RL, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage* 2007;36:630–44.
- [21] Hua K, Zhang J, Wakana S, Jiang H, Li X, Reich DS, et al. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *Neuroimage* 2008;39:336–47.
- [22] Archer DB, Vaillancourt DE, Coombes SA. A template and probabilistic atlas of the human sensorimotor tracts using diffusion MRI. *Cereb Cortex* 2018;28:1685–99.
- [23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
- [24] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. 2nd ed New York: Springer series in statistics; 2001.
- [25] Sánchez-Rico M, Alvarado JM. A machine learning approach for studying the comorbidities of complex diagnoses. *Behav Sci* 2019;9:122.
- [26] Goodkin DE, Cookfair D, Wende K, Bourdette D, Pulicino P, Scherokman B, et al. Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke expanded disability status scale (EDSS). *Neurology* 1992;42:859.

9.2 DL-Generic: A deep-learning framework dedicated to medical imaging analyse

In line with my PhD activities, I developed an ‘in-house’ python deep-learning framework called ‘dl-generic’. This project was initially designed to allow people with no IT background to be able to use classical deep-learning models for their own studies. This could concern biologists from the lab who are interested in new DL techniques but often lack the technical knowledge to actually develop the associated tools. In consequence, the framework had to be user-friendly and to propose at least one model for each type of ‘classical’ DL task (regression, classification, generation or segmentation). It also had to be collaborative, in a sense that every user should be able to develop their own specific models (in coherence to the global architecture) and make them available to other users afterwards.

To begin with, I developed classical CNN and U-Net models, which are very popular for medical analysis. It allows users to run regression, segmentation or image generation tasks. These models were tested on dedicated tasks. For CNN, I tried to predict the age of a subject based on a T1w acquisition using the ABIDE database. The segmentation part (U-Net) was evaluated on a brain mask task. The generating capacity of the U-Net was tested on a zero-Gado study in which we tried to predict the effect of gadolinium on a T1w sequence.

Regarding image generation, I quickly upgraded the U-Net model to a GAN version by adding a CNN-discriminator into the model. As in classical adversarial networks, the CNN tries to predict if a given input is generated by the U-Net or is actually real. On the other side, the U-Net, by integrating CNN’s loss into its own one, will learn how to fool the CNN. This procedure is used in most state-of-the-art generating models. It forces the generator to produce more realistic outputs.

Further on, during my investigation into data-harmonization, I developed a cycleGAN (subsection 3.2.3 model implementation). The model was used to translate images style between two domains (in my case sites), using two complementary GANs. During all my research in MR harmonization more complex models have been included in the project, like VAE, VAE-GAN and ImUnity.

This project is in open-access here: https://github.com/nifm-gin/dl_generic.

The usage of all models has been organized as follows:

1. Dispatch the data into three folders train/val/test, each using the BIDS nomenclature.
2. Generate patches from the data. This is done with the script ‘patches_generation.py’. This step is crucial as every further step will work on those generated patches. The user can generate patches of the desired shapes (either in 2D or in 3D). Additional optional arguments can be specified for a more specific usage, like for example if the user wants some overlap between the patches, or to use 2D patches for a 2.5D approach.
3. Use the model of user’s choice on the patches previously generated. This is done with the script ‘use_model.py’. It can be used to train the model and run inference or just one of these two steps. For this step the user can specify every training parameter needed (learning rate, batch-size, optimizer, number of layers, the loss etc...).
4. In the case of a segmentation or generation task, the user can reconstruct the model’s outputs. In fact, as working on patches, the outputs are based on input patches so a reconstruction step has to occur to have the final output. This is done with the script ‘reconstruction.py’, which will do the opposite of the patch generation script.

Every model can be used in a 2D, 2.5D (segmentation and generation models only) or 3D fashion. The user can also specify if training should be run on GPU or CPU and he/she can also specify the number of GPU to use. A visualization of training is also possible via the integration of Tensorboard. Users can also stop any training and resume it later on as model checkpoints are saved regularly during training.

So far, the framework has been mainly used by 3 students with different backgrounds. 1) Loïc Legris, a neuro-radiologist used it during his medicine thesis. He was focusing on the diagnosis of hemorrhagic transformation for patients in patients with stroke. His project was to use DL techniques to compare the obtained results to the literature. He had no IT background which made the task challenging. He used `dl_generic` for his studies which allowed him to implement and train a CNN for his classification task. He was able to handle the framework very quickly and became autonomous quickly. 2) Yunshi Han, an engineering student who did her ‘end of study’ project at the GIN in 2021. She worked on multi-classes segmentation focusing on Glioma data. As she had a very good IT knowledge, she became autonomous very quickly. She first used the classical U-Net developed for classical one-class segmentation and then upgraded it to a multi-class version. Additionally, she also worked on transfer-learning concerning medical segmentation tasks. These two options were added to the `dl_generic` project. 3) Finally Constance Sohler, an engineering student doing her ‘end of study’ project under my supervision, used `dl_generic` for her project. As she was focusing on MR harmonization, she used the cycleGAN and ImUnity implementations. Having no previous experience in DL, she was able to handle the tools very quickly and helped me by correcting several issues in the code.

At the moment, there is a research engineer from CREATIS in Lyon who is using ImUnity for her project and would like to observe the effect of harmonization on her data. I am very curious to see the results as she is working on pelvis MRI data. Another team from Chapel Hill in the US has requested access to our code to use ImUnity in their studies.

9.3 Colaboration with K  vin Yaury

During my PhD project I was also in contact with other members of the multiomic MIAI chair. This includes PhD students involved in different medical fields, like genomic or proteomic. Once in a month, we met to exchange about our personal project and present our latest advances. It was a great opportunity to get external views on our work and to get relevant recommendations from specialists from different fields.

In particular, I had several exchanges with Kevin Yaury, a Medical Geneticist PhD student who worked with SeqOne under the supervision of Dr Julien Thevenon. As a geneticist, he gave me more insights into the main challenges of the DEFIDIAG project and how we could process the data efficiently. On the other hand, as a data science engineer, I helped him for his needs for data visualization or dimensional reduction. As a result, I contributed to the paper entitled ‘Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine’ (Yaury et al., 2022) which has been published this year. You will find below the open access version of the manuscript.



ARTICLE

Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine



Kevin Yauy^{1,2,*} , François Lecoquierre³, Stéphanie Baert-Desurmont³, Detlef Trost⁴, Aicha Boughalem⁴, Armelle Luscan⁴, Jean-Marc Costa⁴, Vanna Geromel⁵, Laure Raymond⁵, Pascale Richard⁶, Sophie Coutant³, Mélanie Broutin², Raphael Lanos², Quentin Fort², Stenzel Cackowski⁷, Quentin Testard^{1,5}, Abdoulaye Diallo², Nicolas Soirat², Jean-Marc Holder², Nicolas Duforet-Frebours², Anne-Laure Bouge², Sacha Beaumeunier², Denis Bertrand², Jerome Audoux², David Genevieve⁸, Laurent Mesnard^{9,10}, Gael Nicolas³, Julien Thevenon¹, Nicolas Philippe²

ARTICLE INFO

Article history:

Received 30 November 2021

Received in revised form

7 February 2022

Accepted 7 February 2022

Available online 17 March 2022

Keywords:

ClinVar

Gene–phenotype associations

Sequencing reinterpretation

Variant pathogenicity

ABSTRACT

Purpose: Retrospective interpretation of sequenced data in light of the current literature is a major concern of the field. Such reinterpretation is manual and both human resources and variable operating procedures are the main bottlenecks.

Methods: Genome Alert! method automatically reports changes with potential clinical significance in variant classification between releases of the ClinVar database. Using ClinVar submissions across time, this method assigns validity category to gene–disease associations.

Results: Between July 2017 and December 2019, the retrospective analysis of ClinVar submissions revealed a monthly median of 1247 changes in variant classification with potential clinical significance and 23 new gene–disease associations. Re-examination of 4929 targeted sequencing files highlighted 45 changes in variant classification, and of these classifications, 89% were expert validated, leading to 4 additional diagnoses. Genome Alert! gene–disease association catalog provided 75 high-confidence associations not available in the OMIM morbid list; of which, 20% became available in OMIM morbid list. For more than 356 negative exome sequencing data that were reannotated for variants in these 75 genes, this elective approach led to a new diagnosis.

Conclusion: Genome Alert! (<https://genomealert.univ-grenoble-alpes.fr/>) enables systematic and reproducible reinterpretation of acquired sequencing data in a clinical routine with limited human resource effect.

© 2022 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Gael Nicolas, Julien Thevenon, and Nicolas Philippe jointly supervised this work.

*Correspondence and requests for materials should be addressed to Kevin Yauy, Institute for Advanced Biosciences, UGA/Inserm U 1209/CNRS UMR 5309 joint research center, Site Santé-Allée des Alpes, 38700 La Tronche, France. E-mail address: kevin.yauy@univ-grenoble-alpes.fr

Affiliations are at the end of the document.

doi: <https://doi.org/10.1016/j.gim.2022.02.008>

1098-3600/© 2022 The Authors. Published by Elsevier Inc. on behalf of American College of Medical Genetics and Genomics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Genetic tests are increasingly prescribed and included in health care pathways for diverse clinical indications.^{1,2} Several countries have developed population genomics organizations that are revolutionizing medical practices.^{3,4} However, many of these genomic analyses remain inconclusive owing to limitations in genomic and medical knowledge available at the time of analysis.

The American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) recommendations for variant classification aim at standardizing variant interpretation practices in genomic centers, in the context of medical interpretation.⁵ Recently, tools have been published to automatically classify genomic variants on the basis of these recommendations.^{6–8} Meanwhile, evolving medical knowledge and rapid adoption of clinical genome sequencing have influenced the standard practices and have created additional needs. A current and major preoccupation in this field is the definition of standards for periodic and prospective reanalysis of existing sequencing data. Indeed, reanalyzing existing genomic data improves diagnostic yield (7% increase per year).^{9,10}

In practice, such an in-depth reinterpretation is mainly manual and time-consuming, with major bottlenecks such as human and funding resources or lack of consistency between centers. Clinical recommendations from the American and European Societies of Human Genetics reinforce the need for a standardized and automated approach to the reinterpretation of genomic analyses.^{11–14} Some companies offer paid black box services, with poorly detailed methods that cannot be reproduced.^{15,16}

Clinical knowledge of rare diseases is contained in expert-curated databases (such as OMIM¹⁷ or Clinical Genome Resource [ClinGen]¹⁸), peer-reviewed medical literature, and information sharing between health practitioners through community-based platforms (such as MatchMaker Exchange¹⁹ or ClinVar²⁰). Reliability and exhaustiveness of information vary widely across these data sources. Furthermore, careful monitoring of clinical knowledge by every laboratory represents an organizational challenge for a prospective reanalysis of acquired data. To enable a systematic, reproducible, and prospective genome interpretation, a collaborative approach for clinical knowledge aggregation combined with automated medical knowledge monitoring and curation is needed.

The main community-based repository of genomic knowledge is ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), a shared variant interpretation database that featured 1 million submissions in 2020. ClinVar is updated weekly with several thousands of modifications of variant classifications that could affect the diagnostic yield of previous analyses. There is currently no monitoring system that can highlight these changes at a scale for the complete database. Besides variant classification, gene–phenotype

association catalogs are crucial because they are commonly used to design phenotype-specific gene panels for dry-lab filtering and set the frontiers for clinical genome analysis.^{21,22} Although not their primary purpose, variant-centered databases could also theoretically provide a complementary resource to gather gene–phenotype knowledge.

In this article, we detail an automated method for the reassessment of variant pathogenicity and gene–phenotype associations through ClinVar follow-up. This procedure, called Genome Alert!, aims at performing a routine and systematic reinterpretation of existing genomic data. The procedure's effectiveness was evaluated through a 29-month multicentric series (2018–2019) of 5959 consecutive individuals screened using targeted sequencing (4929 individuals with hereditary cancers) and exome sequencing (1000 analyses including 356 undiagnosed individuals with suspected Mendelian disorders).

Materials and Methods

Genome Alert! standardized procedure

ClinVCF, Variant Alert!, and ClinVarome are a suite of tools that constitute the heart of the Genome Alert! standardized procedure.

ClinVCF: A ClinVar quality processing method

Before comparing different versions of the same source, data consistency needs to be verified. This first step is based on ClinVCF tool, and once every submission has been tracked, data will be processed for the next step.

ClinVCF imports monthly updated ClinVar Xtensible Markup Language (XML) files. XML format was preferred over VCF mainly because of better consistency and traceability across versions for the ClinVar Variation ID, the history of changes in each variant classification, and the additional gene–phenotype data availability in XML. ClinVCF considers an automatic reclassification of variants with at least 4 submissions and conflicting interpretations of pathogenicity status. Consensus classification according to ClinVar policies sets the conflicting interpretations of pathogenicity status when at least 1 conflict in submission is observed, except if an expert consortium (as ClinGen) has defined classification (details available in [Supplemental Method 1](#)). On the basis of the provided classifications transformed from literal transcription (eg, likely pathogenic) to class number (eg, class 4), if ≥ 4 submissions are available, a new consensus is proposed after outlier submissions removal according to the 1.5* Interquartile Range (IQR) Tukey method.²³ We only reclassify variants from conflicting status to likely pathogenic or pathogenic status. ClinVCF provides a 3-tier reclassification confidence score detailed in [Supplemental Figure 1](#). As an output, ClinVCF writes a Variant Calling File (VCF) v4.2 file.

Variant Alert!: A variant knowledge monitoring tool

Variant Alert! tool aims at identifying changes in variant classification across 2 versions of the database. Changes were defined as (1) a modification in the classification of an existing variant and (2) the creation or suppression of a variant entry.

Stratification of the consequences in classification modification was proposed (Supplemental Table 1). Major classification modification was defined as a change that may affect the clinical management of a patient (eg, uncertain significance to likely pathogenic status). Minor classification modification was defined as a change that may not affect the clinical management of a patient (eg, pathogenic to likely pathogenic status).

Variant Alert! writes 2 files: (1) the list of variants that were modified, added, or removed and (2) the list of genes that were added to or removed from the database. This gene list is notably used by ClinVarome.

ClinVarome: A method for automated gene–disease association evaluation

ClinVarome tool aims to periodically and automatically evaluate gene–disease association in the ClinVar database. To differentiate genes on the basis of their clinical validity, the work from European Molecular Biology Laboratory–European Bioinformatics Institute Gene2Phenotype,²⁴ ClinGen,¹⁸ and Genomic England PanelApp²⁵ were first compared. Although theoretically comparable, their rationales and contents were partially overlapping and with conflicting classifications. To discriminate candidate genes from definitive gene–disease associations, we decided to use an unsupervised clustering model. Only the genes with at least 1 likely pathogenic or pathogenic variant (single nucleotide variant or indel affecting a single gene) in ClinVar were considered in a list called ClinVarome. As a consensus criterion, we chose to assess the strength of a gene–disease association through the quantification of 4 variables: (1) count of likely pathogenic and pathogenic variants, (2) highest variant classification (CLNSIG, likely pathogenic or pathogenic), (3) highest ClinVar review variant confidence (CLNREVSTAT, from 0 to 4 stars), and (4) time interval between the first and the last pathogenic variant submission (replication of the gene–disease association event). For these 4 variables, values were gathered through periodic monitoring of changes in the database following the ClinVCF and Variant Alert! tool procedures. Clustering variants according to these variables allowed us to define clusters of genes according to their clinical validity. The scikit-learn Agglomerative Clustering tool (parameters: Euclidean affinity, ward linkage) was used, and t-distributed stochastic neighbor embedding representation (parameters: 2 components, perplexity 150, 2000 iterations, and 1000 iterations without progress) was performed. Gene–disease validity classification was computed per gene but not per disease. The Gene Curation Coalition (GenCC) (<https://thegencc.org/>) database was released recently and was used to evaluate ClinVarome. To compare ClinVarome

clusters and GenCC classification, GenCC submissions were summarized into 3 categories (Green, Orange, Red) (Supplemental Methods 2).

Study design and participants

To evaluate the clinical impact of Genome Alert!, we collected 5929 consecutive germline sequencing data samples from 3 centers in France between July 2017 and December 2019 as part of their routine genetic investigation: (1) a variant database gathering all class 3 (uncertain significance), class 4 (likely pathogenic), and class 5 (pathogenic) variants identified in a colon cancer–targeted sequencing (14 genes) sequenced in 2540 individuals in the Rouen University Hospital; (2) a cancer-targeted sequencing data set of 2389 individuals by the Cerba laboratory (66 genes); and (3) exome sequencing data of individuals with developmental disorders, rare kidney diseases, or other rare diseases as follows: 108 probands from the Rouen University Hospital, 477 probands (with 356 negative analysis) from the Cerba laboratory, and 415 probands from the Eurofins Biomnis laboratory. Patient samples, together with a basic phenotype description and molecular diagnosis (when available), were anonymized. Two main clinical evaluations were performed: (1) variant-centered reanalysis, which aims at matching individuals that carry exact variants with potential clinical significance reported by Genome Alert!, and (2) gene-centered reanalysis, which aims at matching individuals who carry candidate variants in high-confidence clinical genes referenced in ClinVarome and not in OMIM. Initial analyses were performed between 0 and 2 years before this reanalysis.

Selection of variants with potential clinical significance

All sequencing data were systematically reinterpreted according to Genome Alert!'s report and compared with the initial variant interpretation. For targeted sequencing and exome reanalysis, genomic positions of variants with major changes in classification were queried in the existing patient's variant calling files (variant-centered analysis). For exome data, we performed a reanalysis of variants in VCF with the following criteria: (1) among 75 ClinVarome morbid genes, which were not available in OMIM, and with a second event of gene–disease validation (including a likely pathogenic or pathogenic variant with ClinVar review confidence ≥ 2 stars and a likely pathogenic or pathogenic variant entry subsequent to the initial entry); (2) variant not shared with another individual in the series; (3) sufficient sequencing quality (variant allele fraction $> 25\%$ and read depth > 20 reads); (4) rare in Genome Aggregation Database²⁶ population (frequency $< 10^{-5}$ if heterozygous genotype or 10^{-4} if homozygous genotype); and (5) protein consequence among nonsense, frameshift, missense (missense are selected with Combined Annotation

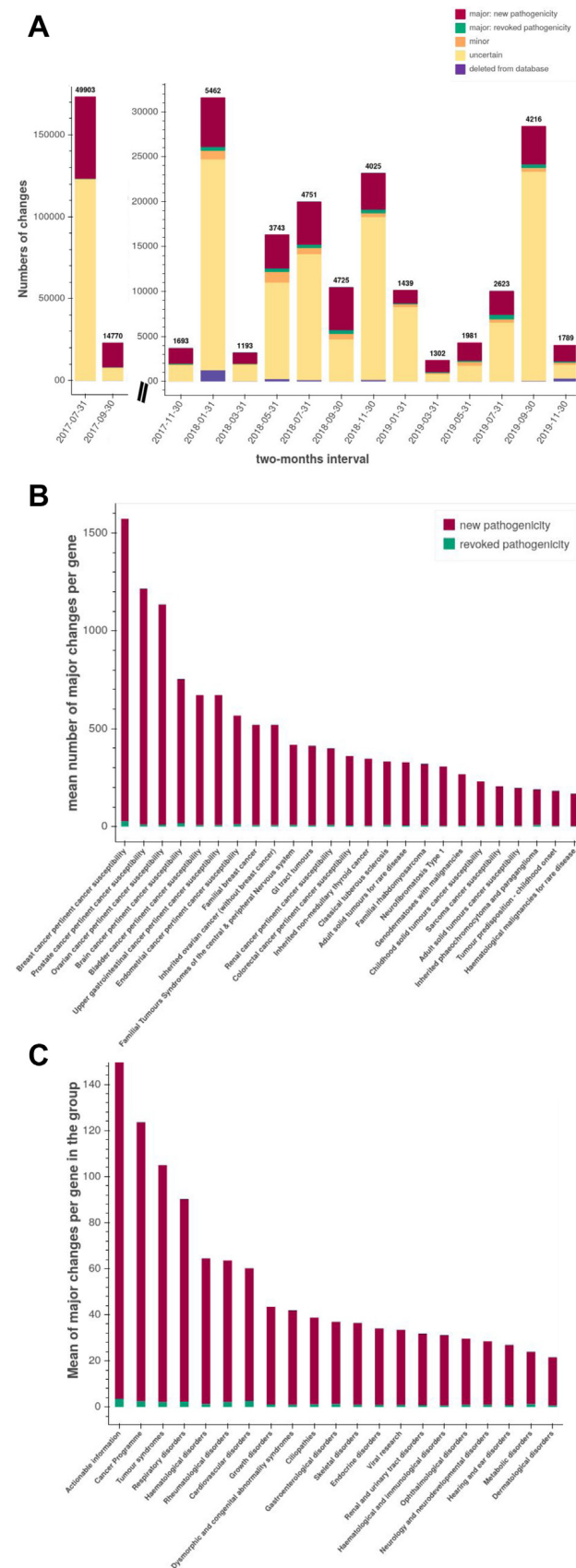


Figure 1 ClinVar variant classification monitoring between July 2017 and December 2019. A. Bar chart distribution of every 2 months of changes in variant classification. The bar chart was

Dependent Depletion²⁷ score > 30 and MetaSVM²⁸ = D), or splice variants (based on dbcsnv RF²⁹ predicted impact score > 0.6) (gene-centered reanalysis).

Results

ClinVar knowledge dynamics

To get insights into variant classification and gene–disease association and to estimate the amount of new clinically relevant information in the ClinVar database available through time, a retrospective analysis of ClinVar submissions over 29 months was performed (July 2017 [included] to December 2019). Of note, VCF genomic positions in ClinVar were introduced in July 2017 and probably are associated with the largest injection in the ClinVar database.

The number of variants with ACMG/AMP classification⁵ increased from 144,943 to 491,838. Among modifications in the database, the count of major changes was 107,167 in ACMG/AMP classification, and among these, 103,615 resulted in a pathogenicity status, which was previously unreported, whereas 3552 resulted in the revocation of a previously established pathogenicity (Figure 1A). These changes varied significantly according to disease groups the between gene panels (according to Genomics England PanelApp), in which the oncogenetic panels were on top of the list of panels. The panels and disease groups presenting most of the changes per gene are presented in Figure 1B and C and Supplemental Table 2. Clinical gene entries in ClinVar were also monitored. A median of 23 ClinVar morbid genes per month that were newly associated with Mendelian disease was observed (Figure 2).

Changes in variant classification

To evaluate the robustness of clinical variant information, the consistency of variant classification was explored and is described in Supplemental Table 3. Among 144,943

split for better readability. Bold numbers and dark red color represent new (likely) pathogenic variant entries, green represents number of revoked (likely) pathogenic variants, orange represents number of minor change variants (eg, pathogenic to likely pathogenic), yellow represents number of changes with uncertain clinical impact (VUS or conflict entry), and purple represents number of changes leading to variant disappearance. B. Bar chart of top panels with clinically significant changes per gene (major changes). Dark red color represents (likely) pathogenic variant entries, and green represents revoked (likely) pathogenic variants. C. Bar chart of top disease group with clinically significant changes per gene (major changes). Dark red color represents (likely) pathogenic variant entries, and green represents revoked (likely) pathogenic variants. GI, gastrointestinal tract; VUS, variant of uncertain significance.

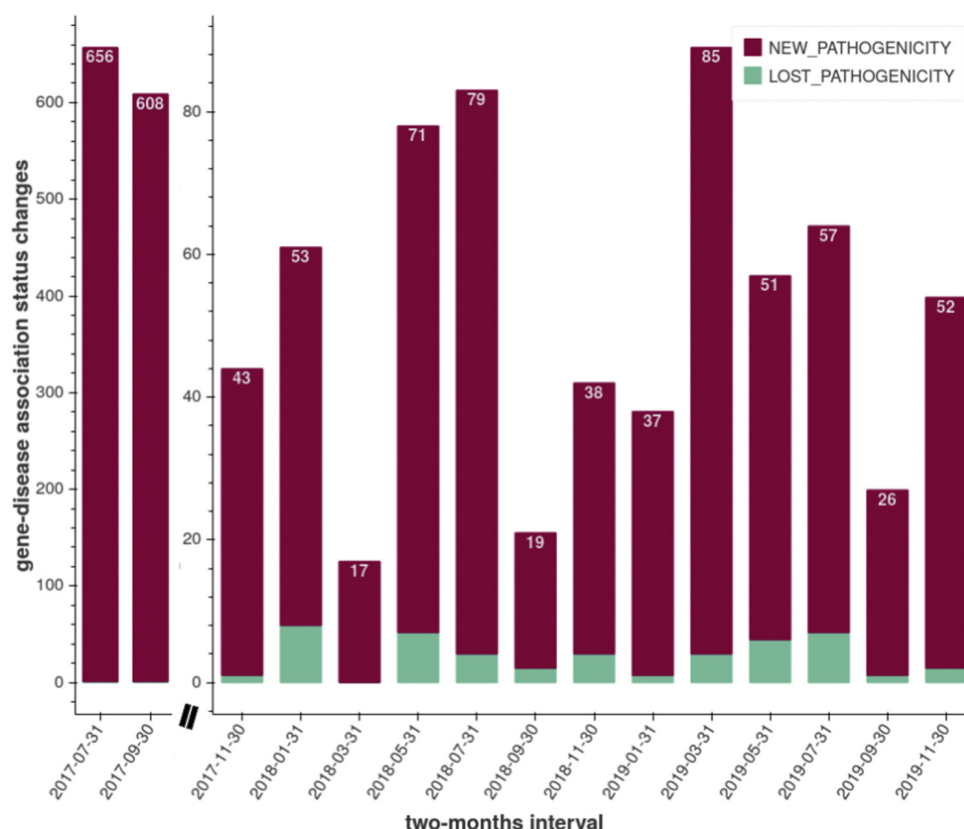


Figure 2 ClinVar clinical genes entries associated with new or deprecated Mendelian disease (morbid status) distribution between December 2017 and December 2019. The bar chart was split for better readability. Dark red represents morbid genes entries (first variant with likely pathogenic or pathogenic status), and green represents revoked morbid genes. White numbers represents number of new morbid gene entries by 2 months.

variants available in July 2017, 10,254 (7%) were reclassified between July 2017 and December 2019, ie, we observed only a small portion of variants being reclassified over time. These reclassifications included automatically reclassified variants with conflicting interpretations. More precisely, among the 11,417 likely pathogenic variants, 1125 (9.94 %) variants were reclassified as benign variants, likely benign variants, variants of uncertain significance, or variants with conflicting interpretations of pathogenicity.

Automatic variant reclassification with conflicting interpretations

A criticism of the ClinVar database is the misclassification of pathogenic variants, such as the well-known *HFE* pathogenic variant NM_000410.3:c.845G>A. We observed that it was mostly due to a unique outlier submission with a classification for a distinct condition (eg, cutaneous photosensitivity porphyria phenotype). We evaluated our method to remove such outlier submissions. Among all the variants available in ClinVar in December 2019, 22,973 of a total of 503,994 (4.5%) variants were classified with a conflicting interpretation of pathogenicity. Genome Alert! automatic reclassification method proposes to detect outlier submissions to suggest a consensus classification. This

allowed the reclassification of 188 variants from conflict to likely pathogenic or pathogenic classification in 135 genes and 1625 variants in 436 genes from conflict to likely benign or benign classification (Supplemental Table 4, Supplemental Figures 1 and 2).

Variants automatically reclassified as likely pathogenic or pathogenic in cancer ($n = 9$) and cardiogenetic disease ($n = 11$) were presented to French National experts in the field. Of these 20 automatic reclassifications, 17 were confirmed as accurate by experts and 3 remained as variants of uncertain significance, lacking evidence of pathogenicity for our experts.

Clinical impact of changes in variant classification

To assess the clinical impact of Genome Alert!'s changes in variant classification, previously analyzed cancer-predisposition targeted sequencing data were assessed (4929 individuals from 2 genetic centers) (variant-centered reanalysis, Figure 3). Among all variants detected in this cohort, this method highlighted 45 variants with major changes between the time of analysis and December 2019, which were proposed for manual review by their referring geneticists (Supplemental Tables 5 and 6).

Among the 45 variants, 30 had been already manually reported by the clinical geneticists as likely pathogenic or

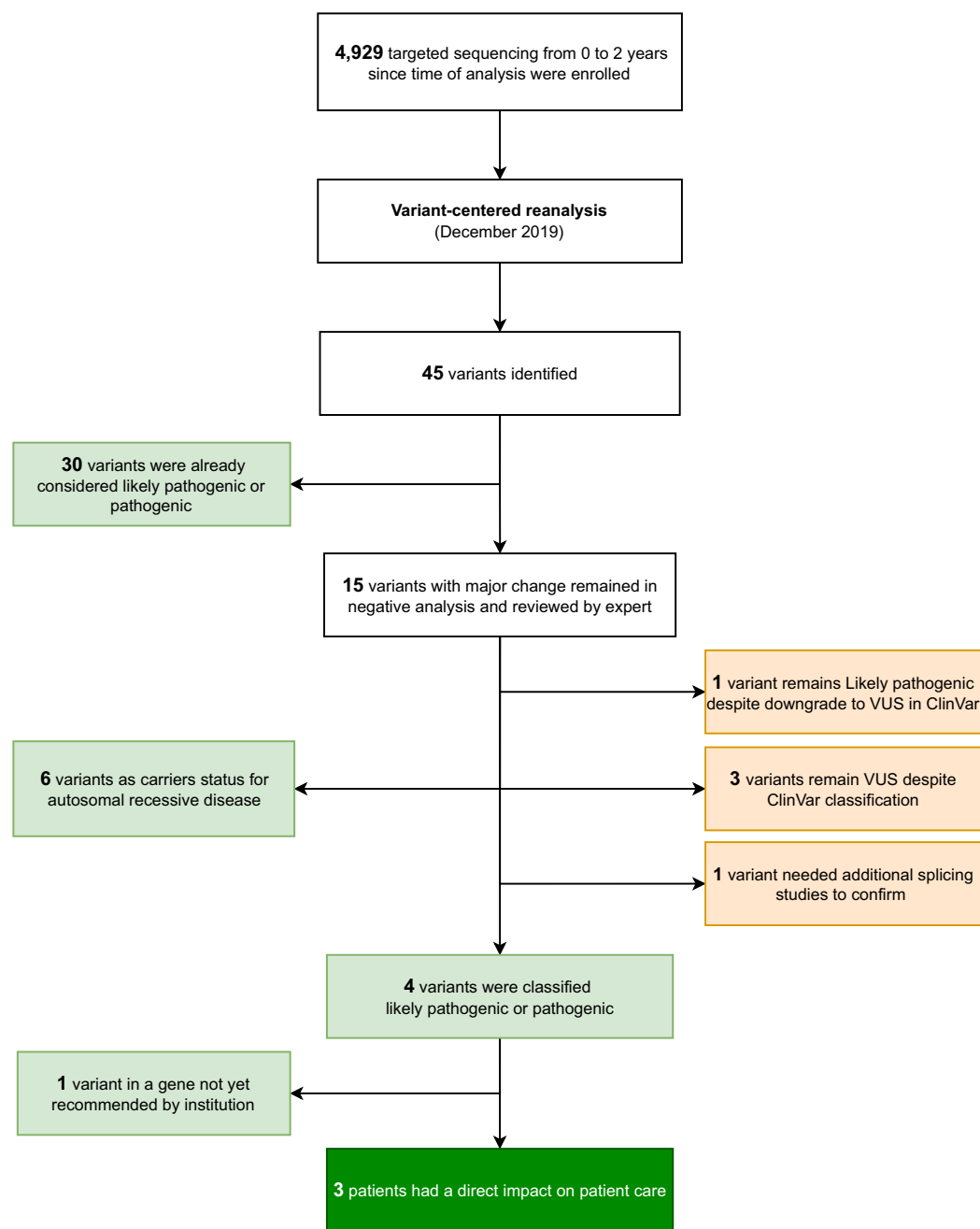


Figure 3 Experimental design of the variant-centered reanalysis. Flow charts describing how the sequencing data were reinterpreted according to variant reclassification only. Green box represents new diagnosis. Light green boxes represent confirmed variant classification. Orange boxes represent excluded variants. VUS, variant of uncertain significance.

pathogenic at the initial time of analysis, meaning that these classifications were ahead of the ClinVar database. The 15 unreported variants were manually curated, looking for additional diagnoses. Among them, 14 variants were newly classified as likely pathogenic or pathogenic and 1 was downgraded as a variant of uncertain significance (VUS) in ClinVar. The manual curation of these 14 variants lead to the conclusion that 6 corresponded to a carrier status for a recessive disorder, 3 were manually classified as VUS, and 5 were submitted to a multidisciplinary meeting for external review. Finally, 4 of these latter 5 were classified as likely

pathogenic or pathogenic by experts leading to additional diagnoses. One variant remained classified as a VUS, and complementary studies on the patient's messenger RNA were proposed before conclusion (*PALB2*, NC_000016.9(NM_024675.3):c.3350+4A>G). Finally, an 89% validation rate (40 of 45) of major changes were observed. This variant reclassification tracking system allowed an additional diagnosis per 1000 analyses.

Replication of the variant-centered reanalysis was performed in the exome sequencing cohort, looking for variant exact match. Selective reanalysis in previous exome

sequencing analysis (1000 individuals in 3 genomic centers) highlighted <1 variant per exome (only 297 variants) with major changes between the time of analysis and December 2019. These 297 variants were then explored by clinical geneticists. Among all 297 variants, 1 variant (*POLG*, NM_002693.2:c.2243G>C) was automatically reclassified as pathogenic by our IQR outlier submission method and was initially reported as VUS, thus helping us to confirm the diagnosis. Compound heterozygosity was observed for a pathogenic variant (*POLG*, NM_002693.3:c.1399G>A). Exome sequencing reanalysis with the variant-centered reanalysis also provides an additional diagnosis per 1000 analyses.

Monitoring ClinVar gene–disease association knowledge

A focus has been toward exploring rarely explored gene–disease association in ClinVar data. To discriminate candidate genes from definitive gene–disease associations in ClinVarome, unsupervised clustering was performed on the basis of the following criteria: (1) count of likely pathogenic and pathogenic variants, (2) highest variant classification, (3) highest ClinVar review variant confidence, and (4) time interval between the first and the last pathogenic variant submission. According to distances between clusters and model dendrogram, the number of clusters was set to 4 (Figure 4). Careful observation of these clusters identified objective patterns to understand the classification. We observed that all genes in the first and second clusters had a reproducibility event (a new likely pathogenic or pathogenic variant entry, the confirmation of the likely pathogenic or pathogenic classification by another submitter or expert panel) in pathogenicity status, thus giving them strong confidence. Genes from the first cluster hold pathogenic variants with ClinVar's ≥ 2 stars of review confidence and the second cluster genes include pathogenic variants with different entry dates and <2 stars of review confidence. Genes in the third cluster had 1 strong argument for pathogenicity but needed another event to be fully confirmed (the third cluster genes contained at least 1 pathogenic variant and all pathogenic entries were added at the same date). Because genes in the fourth cluster were only likely pathogenic variants, their gene–disease association remained to be confirmed (Supplemental Table 7).

To assess the exhaustivity of the ClinVarome, a comparison with the OMIM database was performed. In December 2019, there was a 95% overlap (3675/3858) between OMIM morbid clinical genes and ClinVarome morbid genes. Overall, 365 genes were referenced only in OMIM and not in ClinVarome. We observed patterns that were not available in ClinVar. These patterns include nonconfirmation of a disorder as a genuine Mendelian disorder (only 1 publication or isolated patient reports), susceptibility to multifactorial disorders or infection, referencing of genes belonging to

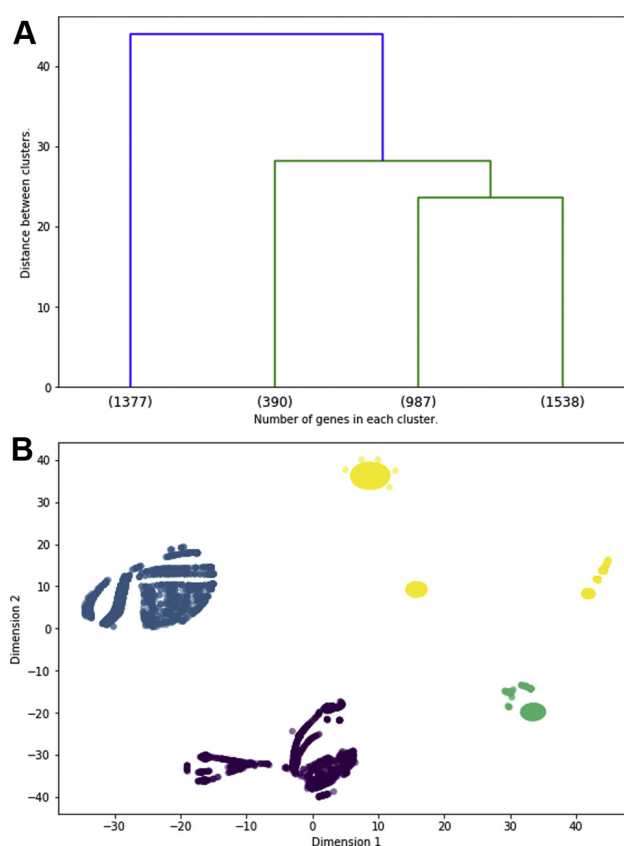


Figure 4 ClinVarome morbid genes exploration and gene–disease validity classification. A. Agglomerative clustering dendrogram of ClinVarome in December 2019. B. t-distributed stochastic neighbor embedding representation of ClinVarome 4 variables by gene data. Green represents fourth cluster (390 genes), yellow represents third cluster (987 genes), blue represents second cluster (1538 genes), and purple represents first cluster (1377 genes).

molecular mechanism distinctive from a single gene disorder as microdeletion or microduplication syndromes, Mendelian traits that are not diseases, epigenetic loci, genes with targeted pathogenic complex variants, and very recently described diseases. The evaluation focused on these 519 specific genes, referenced only in ClinVar and not in OMIM, to assess their potential value in additional diagnoses.

Among the 519 ClinVarome only genes in December 2019, 15 genes were in the first cluster, 60 genes were in the second cluster (ie, 75 high-confidence genes), 140 genes were in the third cluster, and 304 genes were in the fourth cluster. Then, we monitored their inclusion in the OMIM morbid list in the upcoming months. Among the 519 genes exclusively referenced in ClinVarome in December 2019, 55 were reported OMIM morbid 8 months later in August 2020, including 15 of the 75 (20%) initial high-confidence genes. Moreover, 125 of the 140 OMIM morbid genes additional entries between December 2019 and August 2020 were also referenced in ClinVarome release of August 2020. This observation suggested that candidate genes in

ClinVarome may be considered as diagnostic genes before the OMIM validation of the gene–disease causality.

Clinical impact of ClinVarome morbid genes not available in OMIM

We evaluated the relevance of this approach by performing a selective reanalysis of a subsample of the new entries in the ClinVarome, focusing only on the 75 genes that were absent from OMIM morbid list and were referenced in ClinVarome's first and second clusters (gene-centered reanalysis). This experiment highlighted 42 variants in 356

negative exome sequencing data. In this data set, 42 variants were prioritized and were proposed for further interpretation. Among them, 39 were excluded by the expert. The experts' arguments included the presence of variants unrelated to the disease phenotype or a single case series available in the literature. A total of 3 variants were further explored with Sanger sequencing validation, of which 2 were excluded because of artifact status or discordant inheritance pattern (Figure 5).

Overall, this method could ascertain a new diagnosis from the 356 negative exome sequencing data. A nonsense *DLG4* variant NM_001128827.1:c.1840C>T was reported

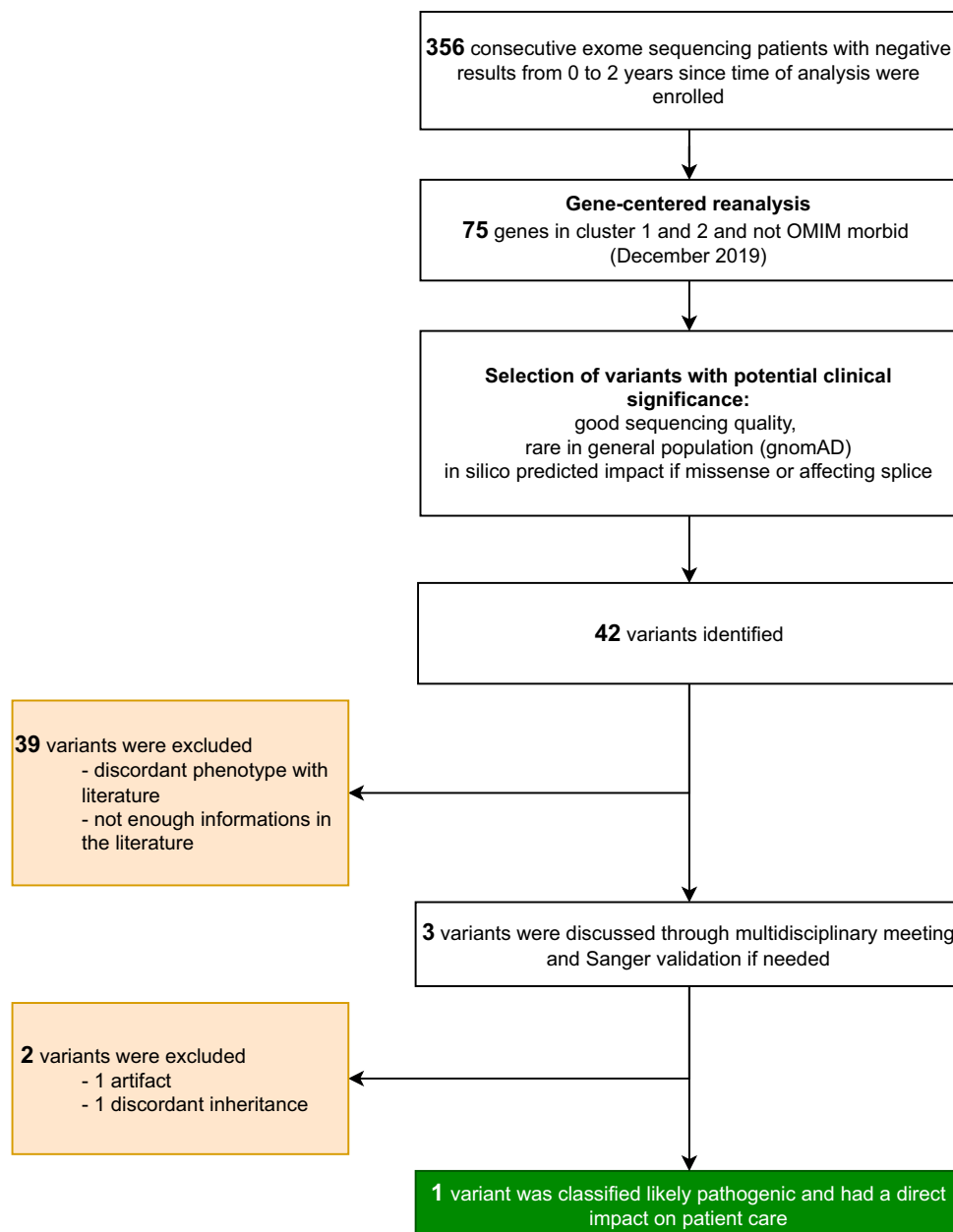


Figure 5 Experimental design for a targeted gene-centered reanalysis. These 75 genes were reported in ClinVarome and not in OMIM and classified as related to a disease (clusters 1 and 2). This list of 75 genes was used for the reinterpretation of negative exome sequencing data ($n = 346$). Green box represents new diagnosis. Orange boxes represent excluded variants. gnomAD, Genome Aggregation Database.

as likely pathogenic, responsible for the patient's phenotype (intellectual disability and microcephaly). Although the first report of *DLG4* association to intellectual developmental disorder was described back in 2016, this gene–disease association was added to the OMIM database only in February 2020.

ClinVarome comparison with the GenCC database

A comparison of gene–disease validity confidence and exhaustivity of ClinVarome with the GenCC database was performed. In October 2021, there was a 65% (3332 of 5187) gene overlap between the 2 databases. Nonoverlapping genes represent mostly the uncertain gene–disease associations from these 2 databases. Exclusive genes in GenCC ($n = 334$) were significantly enriched in orange and red genes (151 of 745 orange genes [$P < .0001$], 158 of 252 red genes [$P < .0001$]). Exclusive genes in ClinVarome ($n = 1471$) were significantly enriched in third and fourth cluster genes (407 of 501 third cluster genes [$P < .0001$], 448 of 743 fourth cluster genes [$P < .0001$]). The 2 databases present a high concordance in gene–disease association confidence (Supplemental Table 8).

Discussion

With the increasing amount of genetic testing performed in health care, there is a critical need for standardized methods to enable prospective genomic data reinterpretation in clinical routine. Through the reassessment of variant pathogenicity and gene–phenotype associations in ClinVar, Genome Alert!'s data mining method proposes the automatic report of a handful of variants that can reasonably be manually interpreted. Our method was applied to a multicentric series of 4929 sequencing tests with various local bioinformatic systems. Genome Alert! successfully allowed new diagnoses in targeted and exome sequencing through query of laboratory's VCFs or variant database and proposed a portable and open-source framework for an automated reanalysis of sequencing data.

Retrospective monitoring of the cutting-edge medical literature on existing genomic data is a major concern for paving the way to genomic medicine.³⁰ There are numerous technical and medical challenges in setting up a routine procedure for reanalysis. This work explored the dynamics of change across all fields of genomic medicine in ClinVar.

Several medical indications for genomic testing were noticed to bear numerous changes in variant classification. Retrospective analysis of the ClinVar database provided an estimation of new clinically relevant information reported each month, which may lead to additional diagnoses in the existing data.³¹ Overall, 9.94 % (1125) of likely pathogenic variants were eventually downgraded and reclassified as benign variants, likely benign variants, variants of uncertain significance, or variants with conflicting interpretation of

pathogenicity in ClinVar over the study period (Supplemental Table 3). This analysis highlights the required carefulness in returning results to the families for likely pathogenic variants because such information could be used for genetic counseling and patient management.

Genome Alert! methods are based on the processing of submissions from the ClinVar full XML release, with no distinction made between submissions with different contexts (eg, somatic or germline status and distinct conditions). Besides, Genome Alert! attributes a unique variant ID on the basis of VCF nomenclature. As such, these variants with potential clinical significance reported by Genome Alert! should be queryable a priori in each genomic center. However, VCF nomenclature is not easy to use with complex variation, which could lead to errors. A switch to the Variation Representation specification from the Global Alliance for Genomics and Health could provide an interesting improvement step.

Clinical effect of changes in variant classification (variant-centered reanalysis) provided in our targeted and exome sequencing cohort provided an additional diagnosis per 1000 analyses. Because time from initial analysis varies from 0 to 2 years, this diagnostic yield will certainly increase with time. This automated system is better for large cohorts of targeted sequencing, with a low number of variants to reinterpret and reaching 10% diagnostic yield in the re-examined variants. Recent literature emphasizes the importance of a standardized procedure adapted for sequencing data reanalysis for considering few candidate variants after an accurate annotation of new gene–phenotype associations and filtering procedure.³⁰

A particular effort was made to evaluate confidence in the reported information to reach a consensus across multiple annotations. The prospective reassessment of ClinVar highlighted numerous conflicts in variant classification. Although our system rarely reclassifies variants with conflicting interpretations, this automatic reclassification method aims to at least remove these potential errors. The expert review of ClinVCF automatic reclassification validates this method on the basis of outlier submission removal using the IQR method, and succeeds in reclassifying abnormalities such as the *HFE* pathogenic variant NM_000410.3:c.845G>A. This work highlights the value of the persistence over time of a classification for relevant genomic information. This work specifically focused on oncogenetics and cardiogenetics, fields in which variant interpretations are particularly conflicting and shifting.^{32,33} Overall, in the ClinVar database, 188 variants could be reclassified in 29 months (ranging from 2017 to 2019). After 8 months, in August 2020, a total of 307 variants were reclassified, highlighting the importance of a systematic and partially automated variant reassessment (Supplemental Figure 2).

Existing literature for gene-centered reanalysis has emphasized the importance of OMIM as an updated resource but not exhaustive.³⁴ To explore and evaluate specifically the ClinVar database for gene-centered reanalysis, we chose to focus our reanalysis on 75 high-confidence ClinVarome

morbid genes (first and second clusters) not available in OMIM morbid genes list. Complementary to OMIM morbid genes, these high-confidence ClinVarome morbid genes from the first and second clusters could provide additional diagnoses in exome or genome sequencing analysis (gene-centered reanalysis). One additional diagnosis was identified with this tight subsampling of variants among the 356 negative exomes, validating the proof of concept. Additional experiments could be performed to fully evaluate the ClinVarome, such as reanalysis with the full list of ClinVarome morbid genes not found in OMIM, additional cohorts, or an extended analysis considering the variants with different phenotypes not reported in the literature.

On the basis of literature data and feature engineering processes from all ClinVarome features during clustering model development, we identified 4 discriminative features for gene–disease clinical validity available in ClinVarome data. Overall, the evaluation relies mainly on the amount of knowledge but also on reported review confidence and more importantly on the time-scale of entries. The Genome Alert! gene-curation via machine learning methods provides an original attempt for automated evaluation of gene confidence in disease. Genome Alert! proposes a standardized clinical validity confidence score that could allow a prospective gene–phenotype association assessment. As such, this approach could be useful to update in silico gene panels. This procedure proposes a complementary approach to the aggregation of multiple expert-reviewed databases such as DDG2P, Genomic England PanelApp, or ClinGen gene–disease validity available in the GenCC database.³⁵ However, ClinVarome gene–disease validity confidence is defined for all diseases associated with a gene, which is less precise than curations submitted to the GenCC database. As ClinVarome is a more exhaustive database, this resource could prioritize genes to be curated by GenCC submitters, particularly in the first and second clusters.

In summary, Genome Alert! highlights changes with potential clinical significance and provides a large retrospective study of a partially automated system for sequencing data reinterpretation. This procedure enables the systematic and reproducible reinterpretation of acquired sequencing data in a clinical routine, with a limited human resource effect and a diagnostic yield improvement. Genome Alert! provides an open-source accessible framework to the community, thus hoping to be applicable in every genetic center.

Data Availability

Software summary

Project name: Genome Alert!

Project home page: <https://genomealert.univ-grenoble-alpes.fr/>

Operating system(s): UNIX (Mac, Linux)

Programming language: Nim, Python, R

License: Apache Licence 2.0

Any restrictions to use by nonacademics: No

Genome Alert! results are publicly available at <https://genomealert.univ-grenoble-alpes.fr/>. Relevant data used to generate Genome Alert! results are available from ClinVar FTP (all monthly ClinVar full XML release data were downloaded from <https://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/>) and in the following resources: OMIM (<https://omim.org/>), Genomic England PanelApp (<https://panelapp.genomicsengland.co.uk/>), and RefSeq annotation (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/GRCh38_latest/refseq_identifiers/GRCh38_latest_genomic.gff.gz). All codes for generating Genome Alert! procedures are available at public GitHub repositories: ClinVCF tool for ClinVar XML full release processing and extraction to VCF format (<https://github.com/SeqOne/clinvcf>), Variant Alert! tool to compare ClinVCF release (https://github.com/SeqOne/variant_alert), ClinVarome tool to evaluate clinical validity of ClinVar morbid genes (<https://github.com/SeqOne/clinvarome>), and the Genome Alert! shiny app (https://github.com/SeqOne/GenomeAlert_app).

Acknowledgments

We sincerely thank all patients, clinicians, biologists, and bioinformaticians involved in this project. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

Author Information

Conceptualization: K.Y., F.L., S.Ca., D.B., A.-L.B., J.A., G.N., J.T., N.P.; Data Curation: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R.; Formal Analysis: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.C., Q.F., J.A.; Funding Acquisition: J.T., N.P.; Methodology: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.Ca., Q.F., J.A.; Project Administration: A.-L.B., J.A., G.N., J.T., N.P.; Resources: K.Y., S.Co., M.B., R.L., Q.F., A.D., N.S., S.B., J.A.; Software: K.Y., S.Co., M.B., R.L., Q.F., A.D., N.S., S.B., J.A.; Supervision: A.-L.B., J.A., G.N., J.T., N.P.; Validation: J.A., G.N., J.T., N.P.; Visualization: A.-L.B., D.B., J.A., D.G., L.M., G.N., J.T., N.P.; Writing-original draft: K.Y., F.L., Q.F., Q.T., J.-M.H., D.B., G.N., J.T., N.P.; Writing-review and editing: K.Y., F.L., S.B.-D., D.T., A.B., A.L., J.-M.C., V.G., L.R., P.R., S.Co., M.B., R.L., Q.F., S.Ca., Q.T., A.D., N.S., J.-M.H., N.D.-F., A.-L.B., S.B., D.B., J.A., D.G., L.M., G.N., J.T., N.P.

Ethics Declaration

Patients referred to the Eurofins Biomnis laboratory, Cerba laboratory, and CHU de Rouen Molecular Genetics laboratory

provided written consent for analysis of their DNA using next-generation sequencing, including research analysis for the purpose of obtaining a molecular diagnosis. Sequencing samples were de-identified. Local Ethics Committee of the CHU Grenoble-Alpes approved the study. Patients or legal guardians provided informed written consent for genetic analyses in a medical setting. This research conforms to the principles of the Helsinki Declaration.

Conflict of Interest

K.Y., M.B., R.L., Q.F., A.D., N.S., D.B., A.-L.B., and N.D.-F. are partially or fully employed by SeqOne Genomics; J.M.-H., S.B., J.A., and N.P. hold shares in SeqOne Genomics; D.T., A.B., A.L., and J.-M.C. are partially or fully employed by Laboratoire Cerba. V.G. and L.R. are partially or fully employed by Laboratoire Eurofins Biomnis. All other authors declare no conflicts of interest.

Additional Information

The online version of this article (<https://doi.org/10.1016/j.gim.2022.02.008>) contains supplementary material, which is available to authorized users.

Affiliations

¹Institute for Advanced Biosciences, Centre de recherche UGA / Inserm U 1209 / CNRS UMR 5309, Grenoble, France; ²SeqOne Genomics, Montpellier, France; ³Department of Genetics and Reference Center for Developmental Disorders, Normandie Center for Genomic and Personalized Medicine, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, F 76000, Rouen, France; ⁴Laboratoire Cerba, Saint-Ouen-l'Aumône, France; ⁵Laboratoire Eurofins Biomnis, Lyon, France; ⁶Unité Fonctionnelle de Cardiogénétique et Myogénétique, Centre de Génétique, Hôpitaux Universitaires Pitié Salpêtrière-Charles Foix, Paris, France; ⁷Grenoble Institut Neurosciences, GIN, Inserm U1216, Université de Grenoble Alpes, Grenoble, France; ⁸Medical Genetic Department for Rare Diseases and Personalized Medicine, Montpellier University Hospital, Montpellier, France; ⁹Soins Intensifs Néphrologiques et Rein Aigu, Hôpital Tenon, Assistance Publique des Hôpitaux de Paris, Paris, France; ¹⁰UMR_S1155, INSERM, Sorbonne Université, Paris, France

References

- Adams DR, Eng CM. Next-generation sequencing to diagnose suspected genetic disorders. *N Engl J Med*. 2018;379(14):1353–1362. <http://doi.org/10.1056/NEJMr1711801>.
- Shendure J, Findlay GM, Snyder MW. Genomic medicine—progress, pitfalls, and promise. *Cell*. 2019;177(1):45–57. <http://doi.org/10.1016/j.cell.2019.02.003>.
- Dollfus H. Le plan France Médecine Génomique 2025 et les maladies rares. *Med Sci (Paris)*. 2018;34(Hors série n°1):39–41. <http://doi.org/10.1051/medsci/201834s121>.
- Turro E, Astle WJ, Megy K, et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*. 2020;583(7814):96–102. <http://doi.org/10.1038/s41586-020-2434-2>.
- Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405–424. <http://doi.org/10.1038/gim.2015.30>.
- Nykamp K, Anderson M, Powers M, et al. Sherlock: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet Med*. 2017;19(10):1105–1117. Published correction appears in *Genet Med*. 2020;22(1):240–242. <https://doi.org/10.1038/gim.2017.37>.
- Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*. 2018;20(9):1054–1060. <http://doi.org/10.1038/gim.2017.210>.
- Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978–1980. <http://doi.org/10.1093/bioinformatics/bty897>.
- Nambot S, Thevenon J, Kuentz P, et al. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet Med*. 2018;20(6):645–654. <http://doi.org/10.1038/gim.2017.162>.
- Wright CF, McRae JF, Clayton S, et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med*. 2018;20(10):1216–1223. <http://doi.org/10.1038/gim.2017.246>.
- Bombard Y, Brothers KB, Fitzgerald-Butt S, et al. The responsibility to recontact research participants after reinterpretation of genetic and genomic research results. *Am J Hum Genet*. 2019;104(4):578–595. <http://doi.org/10.1016/j.ajhg.2019.02.025>.
- Clayton EW, Appelbaum PS, Chung WK, Marchant GE, Roberts JL, Evans BJ. Does the law require reinterpretation and return of revised genomic results? *Genet Med*. 2021;23(5):833–836. <http://doi.org/10.1038/s41436-020-01065-x>.
- Carrieri D, Howard HC, Benjamin C, et al. Recontacting patients in clinical genetics services: recommendations of the European Society of Human Genetics. *Eur J Hum Genet*. 2019;27(2):169–182. <http://doi.org/10.1038/s41431-018-0285-1>.
- Deignan JL, Chung WK, Kearney HM, et al. Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2019;21(6):1267–1270. <http://doi.org/10.1038/s41436-019-0478-1>.
- Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing data. *N Engl J Med*. 2019;380(25):2478–2480. <http://doi.org/10.1056/NEJMc1812033>.
- James KN, Clark MM, Camp B, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *NPJ Genom Med*. 2020;5:33. <http://doi.org/10.1038/s41525-020-00140-1>.
- Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(Database issue):D789–D798. <http://doi.org/10.1093/nar/gku1205>.
- Rehm HL, Berg JS, Brooks LD, et al. ClinGen—the clinical genome resource. *N Engl J Med*. 2015;372(23):2235–2242. <http://doi.org/10.1056/NEJMs1406261>.
- Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36(10):915–921. <http://doi.org/10.1002/humu.22858>.

20. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(Database issue):D980–D985. <http://doi.org/10.1093/nar/gkt1113>.
21. Tumienė B, Maver A, Writzl K, et al. Diagnostic exome sequencing of syndromic epilepsy patients in clinical practice. *Clin Genet.* 2018;93(5):1057–1062. <http://doi.org/10.1111/cge.13203>.
22. Pengelly RJ, Ward D, Hunt D, Mattocks C, Ennis S. Comparison of Mendeliome exome capture kits for use in clinical diagnostics. *Sci Rep.* 2020;10(1):3235. <http://doi.org/10.1038/s41598-020-60215-y>.
23. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2011;1(1):73–79. <http://doi.org/10.1002/widm.2>.
24. Thormann A, Halachev M, McLaren W, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat Commun.* 2019;10(1):2373. <http://doi.org/10.1038/s41467-019-10016-3>.
25. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet.* 2019;51(11):1560–1565. <http://doi.org/10.1038/s41588-019-0528-2>.
26. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434–443. Published correction appears in *Nature.* 2021;590(7846):E53. Published correction appears in *Nature.* 2021;597(7874):E3–E4. <https://doi.org/10.1038/s41586-020-2308-7>.
27. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2019;47(D1):D886–D894. <http://doi.org/10.1093/nar/gky1016>.
28. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* 2020;12(1):103. <http://doi.org/10.1186/s13073-020-00803-9>.
29. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* 2014;42(22):13534–13544. <http://doi.org/10.1093/nar/gku1206>.
30. Matalonga L, Hernandez-Ferrer C, Piscia D, et al. Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *Eur J Hum Genet.* 2021;29(9):1337–1347. Published correction appears in *Eur J Hum Genet.* 2021;29(9):1466–1469. <https://doi.org/10.1038/s41431-021-00852-7>.
31. Landrum MJ, Kattman BL. ClinVar at five years: delivering on the promise. *Hum Mutat.* 2018;39(11):1623–1630. <http://doi.org/10.1002/humu.23641>.
32. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med.* 2016;375(7):655–665. <http://doi.org/10.1056/NEJMsa1507092>.
33. Li D, Shi Y, Li A, et al. Retrospective reinterpretation and reclassification of BRCA1/2 variants from Chinese population. *Breast Cancer.* 2020;27(6):1158–1167. <http://doi.org/10.1007/s12282-020-01119-7>.
34. Bruel AL, Nambot S, Quéré V, et al. Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. *Eur J Hum Genet.* 2019;27(10):1519–1531. <http://doi.org/10.1038/s41431-019-0442-1>.
35. Lazo de la Vega L, Yu W, Machini K, et al. A framework for automated gene selection in genomic applications. *Genet Med.* 2021;23(10):1993–1997. <http://doi.org/10.1038/s41436-021-01213-x>.

9.4 Teaching & supervising experience

During these three years, I had the opportunity to teach at the Grenoble INP Ensimag and Phelma. My first teaching experience was at Ensimag in January 2020 where I supervised a class of 30 students working on a one month project. This project was a 'software engineer project' in which groups of 5 students had to develop a compiler for a programming language close to Java. The project was composed of lectures to introduce the project and the main challenges, then the students would work in groups by themselves and would have some practical sessions to check their progress. Each group had also three intermediate presentations before the final rendering. During these ones, I was in charge of evaluating their progress and alerting them if they had fallen too far behind on the overall project. Finally I had to evaluate their compiler by testing it on a large testing database developed by my colleagues.

This course counted for 55 hours of teaching and I did it again in January 2021. I chose not to do it in 2022 because of a busy schedule at this time, this project having been my main occupation for a month the previous years.

Over April and May 2021, I supervised C programming practical sessions at Phelma. The students were in their first year of engineering and had taken an introductory C programming course before. I gave 14 hours of practical sessions.

In September 2022 I was recruited again to supervise C programming practical sessions, as for 2021 this contract was for 14 hours as well.

My last teaching experience was during the AI4Health winter school in January 2022. With some colleagues from the GIN, we organized a practical session focusing on Deep generating models for gadolinium contrast generation based on T1w acquisitions. We introduced the concept of GAN, transfer learning and contrastive learning. This course lasted 8 hours and was given twice.

In overall I gave 138 hours of teaching, I really enjoyed these teaching experiences. This gave me the opportunity to discover the teaching in different contexts.

Bibliography

- Acquitter, C., Piram, L., Sabatini, U., Gilhodes, J., Moyal Cohen-Jonathan, E., Ken, S., Lemasson, B., 2022. Radiomics-Based Detection of Radionecrosis Using Harmonized Multiparametric MRI. *Cancers* 14, 286. URL: <https://www.mdpi.com/2072-6694/14/2/286>, doi:[10.3390/cancers14020286](https://doi.org/10.3390/cancers14020286). number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- Bashyam, V.M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C.L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Satterthwaite, T.D., Wolf, D.H., Gur, R.E., Gur, R.C., Morris, J.C., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, N.R., Wittfeld, K., Bülow, R., Wolk, D.A., Shou, H., Nasrallah, I.M., Davatzikos, C., The iSTAGING and PHENOM consortia, 2021. Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors. *Journal of Magnetic Resonance Imaging* n/a. URL: <http://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.27908>, doi:[10.1002/jmri.27908](https://doi.org/10.1002/jmri.27908). number: n/a _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.27908>.
- Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., Linn, K.A., 2020. Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage* 220, 117129. URL: <http://www.sciencedirect.com/science/article/pii/S1053811920306157>, doi:[10.1016/j.neuroimage.2020.117129](https://doi.org/10.1016/j.neuroimage.2020.117129).
- Bell, T.K., Godfrey, K.J., Ware, A.L., Yeates, K.O., Harris, A.D., 2022. Harmonization of multi-site MRS data with ComBat. *NeuroImage* 257, 119330. URL: <https://www.sciencedirect.com/science/article/pii/S1053811922004499>, doi:[10.1016/j.neuroimage.2022.119330](https://doi.org/10.1016/j.neuroimage.2022.119330).
- Binquet, C., Lejeune, C., Faivre, L., Bouctot, M., Asensio, M.L., Simon, A., Deleuze, J.F., Boland, A., Guillemin, F., Seror, V., Delmas, C., Espérou, H., Duffourd, Y., Lyonnet, S., Odent, S., Heron, D., Sanlaville, D., Frebourg, T., Gerard, B., Dollfus, H., 2022. Genome Sequencing for Genetics Diagnosis of Patients With Intellectual Disability: The DEFIDIAG Study. *Frontiers in Genetics* 12. URL: <https://www.frontiersin.org/articles/10.3389/fgene.2021.766964>.
- Bottani, S., Burgos, N., Maire, A., Wild, A., Ströer, S., Dormont, D., Colliot, O., 2022. Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. *Medical Image Analysis* 75, 102219. URL: <https://www.sciencedirect.com/science/article/pii/S1361841521002644>, doi:[10.1016/j.media.2021.102219](https://doi.org/10.1016/j.media.2021.102219).

- Brain Development Cooperative Group, 2012. Total and regional brain volumes in a population-based normative sample from 4 to 18 years: the NIH MRI Study of Normal Brain Development. *Cerebral Cortex* (New York, N.Y.: 1991) 22, 1–12. doi:[10.1093/cercor/bhr018](https://doi.org/10.1093/cercor/bhr018). number: 1.
- Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. Technical Report arXiv:1711.09020. arXiv. URL: <http://arxiv.org/abs/1711.09020>, doi:[10.48550/arXiv.1711.09020](https://doi.org/10.48550/arXiv.1711.09020). issue: arXiv:1711.09020 Issue: arXiv:1711.09020 arXiv:1711.09020 [cs] type: article.
- Christensen, J.D., 2003. Normalization of brain magnetic resonance images using histogram even-order derivative analysis. *Magnetic Resonance Imaging* 21, 817–820. doi:[10.1016/s0730-725x\(03\)00102-4](https://doi.org/10.1016/s0730-725x(03)00102-4). number: 7.
- Da-ano, R., Masson, I., Lucia, F., Doré, M., Robin, P., Alfieri, J., Rousseau, C., Mervoyer, A., Reinhold, C., Castelli, J., De Crevoisier, R., Rameé, J.F., Pradier, O., Schick, U., Visvikis, D., Hatt, M., 2020. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports* 10, 10248. URL: <http://www.nature.com/articles/s41598-020-66110-w>, doi:[10.1038/s41598-020-66110-w](https://doi.org/10.1038/s41598-020-66110-w). number: 1.
- Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980. doi:[10.1016/j.neuroimage.2006.01.021](https://doi.org/10.1016/j.neuroimage.2006.01.021). number: 3.
- Dewey, B.E., Zhao, C., Reinhold, J.C., Carass, A., Fitzgerald, K.C., Sotirchos, E.S., Saidha, S., Oh, J., Pham, D.L., Calabresi, P.A., van Zijl, P.C.M., Prince, J.L., 2019. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging* doi:[10.1016/j.mri.2019.05.041](https://doi.org/10.1016/j.mri.2019.05.041).
- Dewey, B.E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E.M., Newsome, S., Oh, J., Calabresi, P.A., Prince, J.L., 2020. A Disentangled Latent Space for Cross-Site MRI Harmonization, in: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racocanu, D., Joskowicz, L. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, Cham. pp. 720–729. doi:[10.1007/978-3-030-59728-3_70](https://doi.org/10.1007/978-3-030-59728-3_70).
- Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.A., Nebel, M.B., Nigg, J.T., O’Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* 19, 659–667. doi:[10.1038/mp.2013.78](https://doi.org/10.1038/mp.2013.78). number: 6.

- Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2020. Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal. *bioRxiv*, 2020.10.09.332973 URL: <https://www.biorxiv.org/content/10.1101/2020.10.09.332973v1>, doi:10.1101/2020.10.09.332973. publisher: Cold Spring Harbor Laboratory Section: New Results.
- Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2021. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage* 228, 117689. URL: <https://www.sciencedirect.com/science/article/pii/S1053811920311745>, doi:10.1016/j.neuroimage.2020.117689.
- Dinsdale, N.K., Jenkinson, M., Namburete, A.I.L., 2022. STAMP: Simultaneous Training and Model Pruning for low data regimes in medical image segmentation. *Medical Image Analysis* 81, 102583. doi:10.1016/j.media.2022.102583.
- Ducharme, S., Albaugh, M.D., Nguyen, T.V., Hudziak, J.J., Mateos-Pérez, J., Labbe, A., Evans, A.C., Karama, S., 2016. Trajectories of cortical thickness maturation in normal brain development — The importance of quality control procedures. *NeuroImage* 125, 267–279. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1053811915009040>, doi:10.1016/j.neuroimage.2015.10.010.
- Eshaghzadeh Torbati, M., Minhas, D.S., Ahmad, G., O'Connor, E.E., Muschelli, J., Laymon, C.M., Yang, Z., Cohen, A.D., Aizenstein, H.J., Klunk, W.E., Christian, B.T., Hwang, S.J., Crainiceanu, C.M., Tudorascu, D.L., 2021. A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. *NeuroImage* 245, 118703. doi:10.1016/j.neuroimage.2021.118703.
- Fischl, B., 2012. FreeSurfer. *NeuroImage* 62, 774–781. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3685476/>, doi:10.1016/j.neuroimage.2012.01.021. number: 2.
- Fortin, J.P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage* 167, 104–120. doi:10.1016/j.neuroimage.2017.11.024.
- Fortin, J.P., Parker, D., Tunc, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., Schultz, R.T., Verma, R., Shinohara, R.T., 2017. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* 161, 149–170. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5736019/>, doi:10.1016/j.neuroimage.2017.08.047.
- Fortin, J.P., Sweeney, E.M., Muschelli, J., Crainiceanu, C.M., Shinohara, R.T., 2016. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* 132, 198–212. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5540379/>, doi:10.1016/j.neuroimage.2016.02.036.
- Gaffney, G.R., Kuperman, S., Tsai, L.Y., Minchin, S., 1989. Forebrain structure in infantile autism. *Journal of the American Academy of Child and Adolescent Psychiatry* 28, 534–537. doi:10.1097/00004583-198907000-00011. number: 4.

- Gagnon-Bartsch, J.A., Speed, T.P., 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13, 539–552. URL: <https://doi.org/10.1093/biostatistics/kxr034>, doi:10.1093/biostatistics/kxr034. number: 3.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-Adversarial Training of Neural Networks. arXiv:1505.07818 [cs, stat] URL: <http://arxiv.org/abs/1505.07818>. arXiv: 1505.07818.
- Giedd, J.N., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., Liu, H., Zijdenbos, A., Paus, T., Evans, A.C., Rapoport, J.L., 1999. Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience* 2, 861–863. doi:10.1038/13158. number: 10.
- Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. arXiv:1406.2661 [cs, stat] URL: <http://arxiv.org/abs/1406.2661>. arXiv: 1406.2661.
- Greve, D.N., Fischl, B., 2018. False positive rates in surface-based anatomical analysis. *NeuroImage* 171, 6–14. URL: <https://www.sciencedirect.com/science/article/pii/S1053811917310960>, doi:10.1016/j.neuroimage.2017.12.072.
- van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.W.L., 2017. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research* 77, e104–e107. doi:10.1158/0008-5472.CAN-17-0339. number: 21.
- Guan, H., Liu, Y., Yang, E., Yap, P.T., Shen, D., Liu, M., 2021. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis* 71, 102076. URL: <https://www.sciencedirect.com/science/article/pii/S1361841521001225>, doi:10.1016/j.media.2021.102076.
- Hagler, D.J., Saygin, A.P., Sereno, M.I., 2006. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage* 33, 1093–1103. doi:10.1016/j.neuroimage.2006.07.036. number: 4.
- Halekoh, U., Højsgaard, S., 2014. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbrktest. *Journal of Statistical Software* 59, 1–32. URL: <https://doi.org/10.18637/jss.v059.i09>, doi:10.18637/jss.v059.i09.
- Hellier, P., 2003. Consistent intensity correction of MR images, in: *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, pp. I–1109. doi:10.1109/ICIP.2003.1247161. iSSN: 1522-4880.
- Huang, X., Belongie, S., 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. Technical Report arXiv:1703.06868. arXiv. URL: <http://arxiv.org/abs/1703.06868>, doi:10.48550/arXiv.1703.06868. issue: arXiv:1703.06868 Issue: arXiv:1703.06868 arXiv:1703.06868 [cs] type: article.

- Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging* 30, 1617–1634. doi:[10.1109/TMI.2011.2138152](https://doi.org/10.1109/TMI.2011.2138152). number: 9.
- Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2018. Image-to-Image Translation with Conditional Adversarial Networks. arXiv:1611.07004 [cs] URL: <http://arxiv.org/abs/1611.07004>.
- Jager, F., Deuerling-Zheng, Y., Frericks, B., Wacker, F., Hornegger, J., 2006. A New Method for MRI Intensity Standardization with Application to Lesion Detection in the Brain. *Vision Modeling and Visualization* , 269–276.
- Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* 5, 143–156. URL: <https://linkinghub.elsevier.com/retrieve/pii/S1361841501000366>, doi:[10.1016/S1361-8415\(01\)00036-6](https://doi.org/10.1016/S1361-8415(01)00036-6). number: 2.
- Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* 8, 118–127. doi:[10.1093/biostatistics/kxj037](https://doi.org/10.1093/biostatistics/kxj037). number: 1.
- Kenward, M.G., Roger, J.H., 1997. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics* 53, 983–997. URL: <https://www.jstor.org/stable/2533558>, doi:[10.2307/2533558](https://doi.org/10.2307/2533558). publisher: [Wiley, International Biometric Society].
- Kingma, D.P., Welling, M., 2014. Auto-Encoding Variational Bayes. URL: <http://arxiv.org/abs/1312.6114>, doi:[10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114). issue: arXiv:1312.6114 arXiv:1312.6114 [cs, stat].
- LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D., 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. Technical Report. URL: <https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1>, doi:[10.1101/2019.12.13.19014902](https://doi.org/10.1101/2019.12.13.19014902). company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press ISSN: 1901-4902 Label: Cold Spring Harbor Laboratory Press Type: article.
- Lange, N., Travers, B.G., Bigler, E.D., Prigge, M.B.D., Froehlich, A.L., Nielsen, J.A., Cariello, A.N., Zielinski, B.A., Anderson, J.S., Fletcher, P.T., Alexander, A.A., Lainhart, J.E., 2015. Longitudinal volumetric brain changes in autism spectrum disorder ages 6-35 years. *Autism Research: Official Journal of the International Society for Autism Research* 8, 82–93. doi:[10.1002/aur.1427](https://doi.org/10.1002/aur.1427). number: 1.
- Langen, M., Schnack, H.G., Nederveen, H., Bos, D., Lahuis, B.E., de Jonge, M.V., van Engeland, H., Durston, S., 2009. Changes in the developmental trajectories of striatum in autism. *Biological Psychiatry* 66, 327–333. doi:[10.1016/j.biopsych.2009.03.017](https://doi.org/10.1016/j.biopsych.2009.03.017). number: 4.
- Leek, J.T., Storey, J.D., 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics* 3, e161. URL: <https://journals.plos.org/plosgenetics/article/doi/10.1371/journal.pgen.0050161>.

- plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030161, doi:10.1371/journal.pgen.0030161. number: 9 Publisher: Public Library of Science.
- Leek, J.T., Storey, J.D., 2008. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* 105, 18718–18723. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.0808709105>, doi:10.1073/pnas.0808709105. number: 48 Publisher: Proceedings of the National Academy of Sciences.
- Lenroot, R.K., Giedd, J.N., 2006. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews* 30, 718–729. doi:10.1016/j.neubiorev.2006.06.001. number: 6.
- Lenroot, R.K., Gogtay, N., Greenstein, D.K., Wells, E.M., Wallace, G.L., Clasen, L.S., Blumenthal, J.D., Lerch, J., Zijdenbos, A.P., Evans, A.C., Thompson, P.M., Giedd, J.N., 2007. Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *NeuroImage* 36, 1065–1073. doi:10.1016/j.neuroimage.2007.03.053. number: 4.
- Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N., 2021. Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Springer International Publishing, Cham. pp. 313–322. doi:10.1007/978-3-030-87199-4_30.
- Liu, Q., Dou, Q., Yu, L., Heng, P.A., 2020. MS-Net: Multi-Site Network for Improving Prostate Segmentation with Heterogeneous MRI Data. arXiv:2002.03366 URL: <http://arxiv.org/abs/2002.03366>.
- Loomes, R., Hull, L., Mandy, W.P.L., 2017. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child and Adolescent Psychiatry* 56, 466–474. doi:10.1016/j.jaac.2017.03.013. number: 6.
- Ma, X., Tan, J., Jiang, L., Wang, X., Cheng, B., Xie, P., Li, Y., Wang, J., Li, S., 2021. Aberrant Structural and Functional Developmental Trajectories in Children With Intellectual Disability. *Frontiers in Psychiatry* 12. URL: <https://www.frontiersin.org/article/10.3389/fpsy.2021.634170>.
- Maaten, L.v.d., Hinton, G., 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605. URL: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>. number: Nov.
- Mahon, R.N., Ghita, M., Hugo, G.D., Weiss, E., 2020. ComBat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Physics in Medicine & Biology* 65, 015010. URL: <https://doi.org/10.1088/2F1361-6560%2F6177>, doi:10.1088/1361-6560/ab6177. number: 1 Publisher: IOP Publishing.
- Modanwal, G., Vellal, A., Buda, M., Mazurowski, M.A., 2020. MRI image harmonization using cycle-consistent generative adversarial network, in: *Medical Imaging 2020: Computer-Aided Diagnosis*, p. 1131413. URL: <https://doi.org/10.1117/1.5300000>.

- [//www.spiedigitallibrary.org/conference-proceedings-of-spie/11314/1131413/MRI-image-harmonization-using-cycle-consistent-generative-adversarial-network/](http://www.spiedigitallibrary.org/conference-proceedings-of-spie/11314/1131413/MRI-image-harmonization-using-cycle-consistent-generative-adversarial-network/) 10.1117/12.2551301.short, doi:10.1117/12.2551301.
- Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J., 2017. Pruning Convolutional Neural Networks for Resource Efficient Inference. URL: <http://arxiv.org/abs/1611.06440>, doi:10.48550/arXiv.1611.06440. arXiv:1611.06440 [cs, stat].
- Nyul, L., Udupa, J., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging 19, 143–150. doi:10.1109/42.836373. number: 2 Conference Name: IEEE Transactions on Medical Imaging.
- Orlhac, F., Frouin, F., Nioche, C., Ayache, N., Buvat, I., 2019. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 291, 53–59. doi:10.1148/radiol.2019182023. number: 1.
- Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I.M., Satterthwaite, T.D., Fan, Y., Launer, L.J., Masters, C.L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S.C., Fripp, J., Koutsouleris, N., Wolf, D.H., Gur, R., Gur, R., Morris, J., Albert, M.S., Grabe, H.J., Resnick, S.M., Bryan, R.N., Wolk, D.A., Shinohara, R.T., Shou, H., Davatzikos, C., 2020. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. NeuroImage 208, 116450. URL: <https://www.sciencedirect.com/science/article/pii/S1053811919310419>, doi:10.1016/j.neuroimage.2019.116450.
- Retico, A., Giuliano, A., Tancredi, R., Cosenza, A., Apicella, F., Narzisi, A., Biagi, L., Tosetti, M., Muratori, F., Calderoni, S., 2016. The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study. Molecular Autism 7, 5. URL: <https://doi.org/10.1186/s13229-015-0067-3>, doi:10.1186/s13229-015-0067-3. number: 1.
- Roca, P., Attie, A., Colas, L., Tucholka, A., Rubini, P., Cackowski, S., Ding, J., Budzik, J.F., Renard, F., Doyle, S., Barbier, E.L., Bousaid, I., Casey, R., Vukusic, S., Lassau, N., Verclytte, S., Cotton, F., Brochet, B., Casey, R., Cotton, F., De Sèze, J., Douek, P., Guillemain, F., Laplaud, D., Lebrun-Frenay, C., Mansuy, L., Moreau, T., Olaiz, J., Pelletier, J., Rigaud-Bully, C., Stankoff, B., Vukusic, S., Marignier, R., Debouverie, M., Edan, G., Ciron, J., Ruet, A., Collongues, N., Lubetzki, C., Vermersch, P., Labauge, P., Defer, G., Cohen, M., Fromont, A., Wiertlewsky, S., Berger, E., Clavelou, P., Audoin, B., Giannesini, C., Gout, O., Thouvenot, E., Heinzlef, O., Al-Khedr, A., Bourre, B., Casez, O., Cabre, P., Montcuquet, A., Créange, A., Camdessanché, J.P., Faure, J., Maurousset, A., Patry, I., Hankiewicz, K., Pottier, C., Maubeuge, N., Labeyrie, C., Nifle, C., Ameli, R., Anxionnat, R., Attie, A., Bannier, E., Barillot, C., Ben Salem, D., Boncoeur-Martel, M.P., Bonneville, F., Boutet, C., Brisset, J.C., Cervenanski, F., Claise, B., Commowick, O., Constans, J.M., Dardel, P., Desal, H., Dousset, V., Durand-Dubief, F., Ferre, J.C., Gerardin, E., Glatard, T., Grand, S., Grenier, T., Guillemin, R., Guttmann, C., Krainik, A., Kremer, S., Lion, S., Menjot de Champfleury, N., Mondot, L., Outteryck, O., Pyatigorskaya, N., Pruvo, J.P., Rabaste, S., Ranjeva, J.P., Roch, J.A., Sadik, J.C., Sappey-Marini, D., Savatovsky, J., Tanguy, J.Y., Tourbah, A., Tourdias, T., 2020. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. Diagnostic

- and Interventional Imaging URL: <http://www.sciencedirect.com/science/article/pii/S2211568420301558>, doi:10.1016/j.diii.2020.05.009.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs] URL: <http://arxiv.org/abs/1505.04597>.
- Sanchez, C.E., Richards, J.E., Almli, C.R., 2012a. Age-specific MRI templates for pediatric neuroimaging. *Developmental Neuropsychology* 37, 379–399. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3399736/>, doi:10.1080/87565641.2012.688900. number: 5.
- Sanchez, C.E., Richards, J.E., Almli, C.R., 2012b. Neurodevelopmental MRI brain templates for children from 2 weeks to 4 years of age. *Developmental psychobiology* 54, 77–91. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3184192/>, doi:10.1002/dev.20579. number: 1.
- Schettini, R., Gasparini, F., Corchs, S., Marini, F., Capra, A., Castorina, A., 2010. Contrast image correction method. *Journal of Electronic Imaging* 19, 023005. URL: <https://www.spiedigitallibrary.org/journals/journal-of-electronic-imaging/volume-19/issue-2/023005/Contrast-image-correction-method/10.1117/1.3386681.full>, doi:10.1117/1.3386681. number: 2 Publisher: SPIE.
- Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T., 2011. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis* 15, 267–282. URL: <https://www.sciencedirect.com/science/article/pii/S1361841510001337>, doi:10.1016/j.media.2010.12.003. number: 2.
- Sherkatghanad, Z., Akhondzadeh, M., Salari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U.R., Khosrowabadi, R., Salari, V., 2019. Automated Detection of Autism Spectrum Disorder Using a Convolutional Neural Network. *Frontiers in Neuroscience* 13, 1325. doi:10.3389/fnins.2019.01325.
- Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., Jarso, S., Pham, D.L., Reich, D.S., Crainiceanu, C.M., 2014. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage : Clinical* 6, 9–19. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215426/>, doi:10.1016/j.nicl.2014.08.008.
- Sled, J., Zijdenbos, A., Evans, A., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging* 17, 87–97. doi:10.1109/42.668698. number: 1 Conference Name: IEEE Transactions on Medical Imaging.
- Stein, C.K., Qu, P., Epstein, J., Burows, A., Rosenthal, A., Crowley, J., Morgan, G., Barlogie, B., 2015. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics* 16, 63. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4355992/>, doi:10.1186/s12859-015-0478-3.
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A.,

- Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* 12, e1001779. URL: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779>, doi:10.1371/journal.pmed.1001779. number: 3 Publisher: Public Library of Science.
- Tanaka, S.C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., Seymour, B., Shimizu, T., Hosomi, K., Saitoh, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Yamashita, O., Kawato, M., Imamizu, H., 2021. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data* 8, 227. URL: <http://www.nature.com/articles/s41597-021-01004-8>, doi:10.1038/s41597-021-01004-8. bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;Neural circuits;Neurological disorders;Psychiatric disorders Subject_term_id: diagnostic-markers;neural-circuit;neurological-disorders;psychiatric-disorders.
- Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging* 29, 1310–1320. doi:10.1109/TMI.2010.2046908. number: 6.
- Vijayakumar, N., Allen, N.B., Youssef, G., Dennison, M., Yücel, M., Simmons, J.G., Whittle, S., 2016. Brain development during adolescence: A mixed-longitudinal investigation of cortical thickness, surface area, and volume. *Human Brain Mapping* 37, 2027–2038. doi:10.1002/hbm.23154. number: 6.
- Voelbel, G.T., Bates, M.E., Buckman, J.F., Pandina, G., Hendren, R.L., 2006. Caudate Nucleus Volume and Cognitive Performance: Are they related in Childhood Psychopathology? *Biological psychiatry* 60, 942–950. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2947855/>, doi:10.1016/j.biopsych.2006.03.071.
- Walsh, M.J.M., Wallace, G.L., Gallegos, S.M., Braden, B.B., 2021. Brain-based sex differences in autism spectrum disorder across the lifespan: A systematic review of structural MRI, fMRI, and DTI findings. *NeuroImage. Clinical* 31, 102719. doi:10.1016/j.nicl.2021.102719.
- Wang, L., Lai, H.M., Barker, G.J., Miller, D.H., Tofts, P.S., 1998. Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. *Magnetic Resonance in Medicine* 39, 322–327. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910390222>, doi:10.1002/mrm.1910390222. number: 2 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.1910390222>.
- Wang, Z., Simoncelli, E., Bovik, A., 2003. Multiscale structural similarity for image quality assessment, in: *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers*, 2003, pp. 1398–1402 Vol.2. doi:10.1109/ACSSC.2003.1292216.
- Weisenfeld, N., Warfteld, S., 2004. Normalization of joint image-intensity statistics in MRI using the Kullback-Leibler divergence, in: *2004 2nd IEEE International Symposium on*

- Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821), pp. 101–104 Vol. 1. doi:[10.1109/ISBI.2004.1398484](https://doi.org/10.1109/ISBI.2004.1398484).
- Welch, B.L., 1947. The generalization of 'student's' problem when several different population variances are involved. *Biometrika* 34, 28–35. URL: <https://academic.oup.com/biomet/article/34/1-2/28/210174>, doi:[10.1093/biomet/34.1-2.28](https://doi.org/10.1093/biomet/34.1-2.28). number: 1-2 Publisher: Oxford Academic.
- Wierenga, L., Langen, M., Ambrosino, S., van Dijk, S., Oranje, B., Durston, S., 2014. Typical development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24. *NeuroImage* 96, 67–72. doi:[10.1016/j.neuroimage.2014.03.072](https://doi.org/10.1016/j.neuroimage.2014.03.072).
- Wilke, M., Schmithorst, V.J., Holland, S.K., 2002. Assessment of spatial normalization of whole-brain magnetic resonance images in children. *Human Brain Mapping* 17, 48–60. doi:[10.1002/hbm.10053](https://doi.org/10.1002/hbm.10053). number: 1.
- Wilke, M., Schmithorst, V.J., Holland, S.K., 2003. Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. *Magnetic Resonance in Medicine* 50, 749–757. doi:[10.1002/mrm.10606](https://doi.org/10.1002/mrm.10606). number: 4.
- Yauy, K., Lecoquierre, F., Baert-Desurmont, S., Trost, D., Boughalem, A., Luscan, A., Costa, J.M., Geromel, V., Raymond, L., Richard, P., Coutant, S., Broutin, M., Lanos, R., Fort, Q., Cackowski, S., Testard, Q., Diallo, A., Soirat, N., Holder, J.M., Duforet-Frebouge, N., Bouge, A.L., Beaumeunier, S., Bertrand, D., Audoux, J., Genevieve, D., Mesnard, L., Nicolas, G., Thevenon, J., Philippe, N., 2022. Genome Alert!: A standardized procedure for genomic variant reinterpretation and automated gene–phenotype reassessment in clinical routine. *Genetics in Medicine* URL: <https://www.sciencedirect.com/science/article/pii/S1098360022006542>, doi:[10.1016/j.gim.2022.02.008](https://doi.org/10.1016/j.gim.2022.02.008).
- Yi, Z., Zhang, H., Tan, P., Gong, M., 2018. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. arXiv:1704.02510 [cs] URL: <http://arxiv.org/abs/1704.02510>. arXiv: 1704.02510.
- Yoon, U., Fonov, V.S., Perusse, D., Evans, A.C., Brain Development Cooperative Group, 2009. The effect of template choice on morphometric analysis of pediatric brain data. *NeuroImage* 45, 769–777. doi:[10.1016/j.neuroimage.2008.12.046](https://doi.org/10.1016/j.neuroimage.2008.12.046). number: 3.
- Zhao, C., Carass, A., Lee, J., He, Y., Prince, J.L., 2017. Whole Brain Segmentation and Labeling from CT Using Synthetic MR Images, in: Wang, Q., Shi, Y., Suk, H.I., Suzuki, K. (Eds.), *Machine Learning in Medical Imaging*, Springer International Publishing, Cham. pp. 291–298. doi:[10.1007/978-3-319-67389-9_34](https://doi.org/10.1007/978-3-319-67389-9_34).
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2018. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv:1703.10593 URL: <http://arxiv.org/abs/1703.10593>.
- Zielinski, B.A., Prigge, M.B.D., Nielsen, J.A., Froehlich, A.L., Abildskov, T.J., Anderson, J.S., Fletcher, P.T., Zygumunt, K.M., Travers, B.G., Lange, N., Alexander, A.L., Bigler, E.D., Lainhart, J.E., 2014. Longitudinal changes in cortical thickness in autism and typical development. *Brain: A Journal of Neurology* 137, 1799–1812. doi:[10.1093/brain/awu083](https://doi.org/10.1093/brain/awu083). number: Pt 6.

- Zuo, L., Dewey, B.E., Carass, A., Liu, Y., He, Y., Calabresi, P.A., Prince, J.L., 2021a. Information-based Disentangled Representation Learning for Unsupervised MR Harmonization. arXiv:2103.13283 [cs, eess] URL: <http://arxiv.org/abs/2103.13283>. arXiv: 2103.13283.
- Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A., 2021b. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage* 243, 118569. URL: <https://www.sciencedirect.com/science/article/pii/S1053811921008429>, doi:10.1016/j.neuroimage.2021.118569.