POLITECNICO DI MILANO

Master's Program in Biomedical Engineering

Department of Electronics, Information and Bioengineering

Harmonising large-scale imaging databases to provide integrated assessments of the role of white matter hyperintensities in cognitive aging

Supervisor: Prof. Giuseppe Baselli

Co-supervisor: Ludovica Griffanti

Eugene Duff

Marcella Laganà

Authors:

Ilaria BERTANI Matr. 883039

Valentina BORDIN Matr. 876512

Academic year 2019-2020

Il ringraziamento più grande va alle nostre famiglie,
che ci hanno accompagnato con grande amore lungo questo cammino.
Un grazie a tutte le persone care,
agli amici di sempre e a quelli trovati negli ultimi anni,
per aver contribuito a rendere quest'avventura unica e indimenticabile.
Grazie al Wellcome Centre for Integrative Neuroimaging di Oxford,
per averci ospitato ed accolto
durante i sei mesi che hanno rappresentato
una delle esperienze più formative della nostra carriera accademica.
Grazie alla Fondazione Don Gnocchi,
per averci seguito con grande partecipazione e coinvolgimento.
Infine, un sentitissimo grazie a Ludovica, Eugene,
Marcella e al Professor Baselli,
senza l'insegnamento e il supporto dei quali,
la realizzazione di questo progetto non sarebbe stata possibile.

Table of Contents

List of Figures

training sets used (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

[4.7] Step 3 – Effect of different Training sets. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

[4.8] Step 4 – Effect of FA exclusion from the training features. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

[4.9] Step 4 – Effect of FA exclusion from the training features. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

[4.10] Step 5 – Effect of the use of a local thresholding method (LOCATE) on BIANCA output. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1

and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

[4.11] Step 5 – Effect of the use of a local thresholding method (LOCATE) on BIANCA output. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

[4.12] Step 6 – Effect of different Training sets. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

[4.13] Step 6 – Effect of different Training sets. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

[4.14] Step 7 – Effect of the local thresholding method (LOCATE). Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj

from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

[4.15] Step 7 – Effect of the local thresholding method (LOCATE). Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

[4.16] Impact of Biasfield correction. Scatter plot relative to the BF (a) and BC case (b), respectively; Importance of the different non-imaging parameters relative to the BF (c) and BC case (d). Bias field correction produces a decrease in importance of the variable “Scanner” (in red) in predicting WMH volume.

[4.17] Impact of different Training sets. Scatter plot relative to the Single Training (a) and Mixed Training (b), respectively; Importance of the different non-imaging parameters for WMH prediction for the Single Training (c) and Mixed Training case (d). The use of a Mixed Training instead of a single one produces a decrease in importance of the variable “Scanner” (in red) in predicting WMH volume.

[4.18] Impact of different Thresholding method. Scatter plot relative to the Global Thresholding (a) and the Local one (LOCATE) (b), respectively; Importance of the different non-imaging parameters for WMH prediction for the Global (c) and Local case (d). The use of a Global Thresholding method produces a decrease in importance of the variable “Scanner” (in red) in predicting WMH volume.

[4.19] Features importance for the prediction of Fazekas score in Whitehall dataset using Elastic Net. This figure has to be compared with Figures 4.16 and 4.17 (c, d).

[4.20] Features importance for the prediction of WMH% in UK Biobank dataset using Elastic Net. Red circles represent the most predictive variables common to both Whitehall and UK Biobank dataset.

[4.21] 3D plot of WMH% distributions with respect to: Age and BMI (a, b), Age and TMT_B (c. d) in Whitehall (b, d) and UK Biobank (a, c). The three surfaces represent respectively: the mean of predictive distribution for each query points (in the middle) and the mean itself +/- standard deviation (Confidence Interval) above and below it. In all the plots the x axis corresponds to "OX.AGE", the z axis corresponds to "WMH%", while the y axis changes every time depending on the variable analysed.

[4.22] Scatterplot of the WMH distribution, obtained by GP trained and tested on UK Biobank, according to age.

[4.23] Scatterplot of the WMH distribution, obtained by GP trained on UK Biobank and tested on Whitehall, according to age.

[4.24] Scatterplot of the WMH distribution, obtained by GP trained on UK Biobank and tested on Whitehall (in yellow) vs the one trained and tested on Whitehall (in red).

[4.25] Analysis of Model residuals plots: (a) residuals versus age for the model trained and tested on Whitehall, (b) residuals versus WMH% for the model trained and tested on Whitehall, (c) residuals versus age for the model trained on UK Biobank and tested on Whitehall, (d) residuals versus WMH% for the model trained on UK Biobank and tested on Whitehall.

[4.26] Bland Altman Plot of GP's model trained and tested on Whitehall versus the one trained on UK Biobank and tested on Whitehall. The blue horizontal line represents the mean difference between methods, while the red ones are drawn at the limits of agreement, defined as the mean difference ± 1.96 SD of differences.

List of Tables

[4.5] Spearman Correlation coefficients between actual and predicted WMH%. (i) model trained on the 12 manually labelled subjects belonging to BB and tested on the whole BB dataset; (ii) model trained on Mixed training 24+24+BB, Rater 2, no FA, local thresholding and (iii) model trained on Mixed training 24+24+BB, Rater 2, no FA, global thresholding, tested on both Whitehall (Whll SC1 and SC2) and UK Biobank (BB). Blue boxes indicate the best performing method for each option.

[4.6] Spearman Correlation coefficients between actual and predicted values of WMH%, obtained by implementing GP using four different combinations of kernels.

Table of Acronyms

Here we present a summary table collecting the most widely used acronyms, that will be repeated several times throughout the chapters of this thesis.

	Acronym
White Matter Hyperintensities	*WMH*
Magnetic Resonance Imaging	*MRI*
Dementia Platform UK	*DPUK*
UK Biobank	*BB*
Whitehall	*Whll*
Scanner 1	*SC1*
Scanner 2	*SC2*
Fluid Attenuated Inversion Recovery	*FLAIR*
Fractional Anisotropy	*FA*
Brain Intensity AbNormality Classification Algorithm	*BIANCA*

k-Nearest Neighbours	*k-NN*
Locally Adaptive Threshold Estimation	*LOCATE*
General Linear Model	*GLM*
Gaussian Process Regression	*GP*
Training 1	*TR1*
Training 2	*TR2*
Rater 1	*R1*
Rater 2	*R2*
Biasfield not removed	*BF*
Biasfield Corrected	*BC*
Dice Similarity Index	*DI*
Cluster-level False Positive Ratio	*Cluster-level FPR*

Abstract

BACKGROUND – The increasing availability of brain imaging data from different studies of aging population offers statistical power and great opportunities to build robust models in age-related pathologies. An important field is the prediction of imaging-derived risk scores for neurodegenerative diseases and cognitive impairment. However, variations in data properties across imaging protocols, used scanner, and populations can severely limit our ability to combine datasets.

White matter hyperintensities (WMHs) are gaining more and more relevance as a marker of potential brain damage in asymptomatic aging, but also in non-aged patients with several neurological and vascular disorders. WMHs are assessed both by MRI and CT. The superior contrast of the former is recognised; however, harmonisation limits are given by the well-known difficulties in scanner-independent MRI calibration.

AIMS – In this context, our project aims to harmonise imaging-derived measures of WMH, across two large DPUK (Dementia Platform UK) datasets: Whitehall (Whll) and UK Biobank (BB). Namely, the percent of WMHs volume vs. the brain volume, WMH%, was considered. Whll represents a multi-centre study gathering data from a single population, acquired with the same acquisition protocol but exploiting two different MRI scanners (SC1 and SC2) to derive the imaging data. BB includes data from a different population, imaged using a third scanner and a different acquisition protocol. For this reason, we divided our work in two separate parts: 1) a retrospective harmonisation across scanners (Whll SC1 vs Whll SC2), added to the pre-existing prospective one, offered by the Whll study design; 2) a fully retrospective harmonisation process, challenging the integration of dataset belonging to significantly heterogeneous populations (Whll and BB).

METHODS – As to imaging data, we exploited an automatic tool (BIANCA), based on k nearest neighbour (k-NN) machine learning, to perform lesion segmentation and we assessed the influence on harmonisation of five main analysis parameters: (i) rater who generated the manual masks used as ground truth for the tool training phase; (ii) biasfield correction of the

RF field inhomogeneities affecting images; (iii) different training dataset used (study specific vs mixed); (iv) Functional Anisotropy (FA) availability; and (v) difference in the thresholding method (global or locally adapted).

On the other hand, for the non-imaging variables, we started harmonising all the ones involved in our study through the creation of a specific pipeline for format conversion. We then created a mathematical model, able to predict the WMH% starting from the integrated non-imaging data. This helped us accounting for: i) the variability related to demographic and clinical characteristics of the individuals; ii) to evaluate the relationship between WMH% and their majors risk factors; and iii) to assess harmonisation on the whole non-annotated cohorts, when the predictive influence of the used scanner was lowered or even negligible.

RESULTS – Firstly, we found a protocol able to harmonise WMH measures across datasets, comprising the following parameters: (i) expert rater to perform the manual labelling phase (ii) biasfield correction of the RF field inhomogeneities (iii) use of a mixed training set, combining information from all of the datasets involved in our analysis (iv) Functional Anisotropy (FA) excluded from the MRI training features and (v) use of global thresholding method (0.9) to binarise results.

Moreover, we managed to implement a specific pipeline (Parser) for the harmonisation of the non-imaging variables involved in our study, that is actually available online on the GitLab Platform. In this context, we fitted an Elastic Net model for WMH% prediction from non-imaging data calibrated on the imaging WMH% derived by the optimal settings we defined. This was a valid support to derive the relative importance of the non-imaging variables, used scanner included. Finally, we tested a Gaussian Process regression of WMH% on the non-imaging data. This non-linear predictor was compared to Elastic Net, as the best performing linear predictor. The resulting performance, in terms of correlation between actual and predicted value, was close to 0.4, comparable with Elastic Net.

CONCLUSION – Our findings attested the existence of a general set of parameters, able to derive comparable WMH% measures across datasets, in the context of automatic lesion segmentation. These results, along with the non-imaging data integration, proved the accomplishment of a robust harmonisation on the different datasets involved in our study, that were finally well combined and compatible. The fair heterogeneity of the addressed

datasets permits to foresee a wider extension of our harmonisation approach to further datasets.

Sommario

CONTESTO – Il recente aumento nella disponibilità dei dati di neuro-imaging, provenienti da diversi studi relativi all'invecchiamento della popolazione, offre una notevole potenza statistica e rappresenta pertanto una buona opportunità per la costruzione di modelli matematici robusti in grado di descrivere le patologie correlate all'età. In questo contesto, un aspetto importante è rappresentato dalla predizione degli score di rischio per le malattie neurologiche e i disturbi cognitivi, derivati a partire dai dati di imaging. Tuttavia, le variazioni nelle proprietà di questi ultimi, dovute a differenze nei protocolli di acquisizione, nei centri clinici di riferimento e nelle popolazioni coinvolte, possono limitare notevolmente la capacità di combinare ed integrare tra loro set di dati diversi.

Le iperintensità della materia bianca (White Matter Hyperintensities, WMH) stanno acquisendo sempre maggiore importanza come indicatori clinici di potenziali danni neurologici, sia nel contesto dell'invecchiamento asintomatico, che in relazione a pazienti che, seppure in giovane età, sono affetti da diversi disturbi neurodegenerativi e vascolari. Le WMH vengono generalmente valutate mediante risonanza magnetica nucleare (MRI) o TAC ma il miglior contrasto del primo rispetto al secondo ha portato alla scelta dell'MRI come tecnica standard per la visualizzazione delle lesioni in questione. Tuttavia, le ben note difficoltà relative al processo di calibrazione delle immagini di risonanza magnetica determinano notevoli limiti nel processo di armonizzazione dei dati acquisiti.

OBIETTIVI – In questo contesto, il nostro progetto mira ad una armonizzazione delle misure di WMH ottenute a partire dai dati di imaging relativi a due grandi dataset DPUK (Dementia Platform UK): Whitehall (Whll) e UK Biobank (BB).

Whll rappresenta uno studio multicentrico che da riferimento ad una singola popolazione, acquisita con lo stesso protocollo ma mediante l'utilizzo di due scanner diversi (SC1 e SC2). BB include, invece, dati provenienti da una popolazione diversa, acquisiti utilizzando un terzo scanner e un protocollo di imaging differente rispetto a quello menzionato in precedenza. Pertanto, il nostro lavoro è stato diviso in due parti distinte: 1) un processo di

armonizzazione retrospettiva tra i due scanner presenti in Whll (Whll SC1 e Whll SC2), che si aggiunge ad una preesistente fase di armonizzazione prospettica, intrinseca allo studio di popolazione in questione; 2) un processo di armonizzazione retrospettiva, che mira all'integrazione di dati appartenenti a popolazioni significativamente eterogenee (Whll vs BB).

Per quanto riguarda i dati di imaging, abbiamo sfruttato un tool automatico (BIANCA) per eseguire la segmentazione delle lesioni di interesse e abbiamo cercato di valutare l'influenza di cinque diversi parametri sulla sua performance finale: (i) rater che ha generato le maschere manuali utilizzate come riferimento per la fase di training (ii) correzione delle disomogeneità nel campo a radiofrequenze (RF) che caratterizza le immagini di risonanza magnetica nucleare (iii) differenza nel gruppo di soggetti utilizzati per il training (study specific/single vs mixed) (iv) presenza della Fractional Anisotropy (FA) tra le features utilizzate e (v) differenza nel metodo di thresholding applicato all'output ottenuto (globale o locale).

Per quanto riguarda le variabili di non imaging, abbiamo cercato di armonizzare tutte quelle coinvolte nella nostra analisi, attraverso la creazione di una specifica pipeline per la conversione dei format. Abbiamo poi creato un modello matematico, in grado di prevedere il volume di WMH a partire dai dati di non-imaging, perfettamente integrati tra loro (pressione sanguigna, BMI, test cognitivi, ecc.). Questo ci ha permesso di prendere in considerazione la variabilità dovuta alle caratteristiche demografiche e cliniche degli individui e, inoltre, a valutare il rapporto tra le WMH e i loro principali fattori di rischio.

RISULTATI – Innanzitutto abbiamo delineato un protocollo in grado di ottenere misure di WMH comparabili tra i diversi dataset a disposizione. Esso si compone di una serie di parametri che vengono di seguito elencati: (i) utilizzo di un rater esperto per la fase di segmentazione manuale (ii) correzione del biasfield presente nelle immagini (iii) uso di un training set misto, che combina informazioni provenienti da tutti i dataset coinvolti nella nostra analisi (iv) Fractional Anisotropy (FA) esclusa dalle features di training e (v) uso di un metodo di thresholding globale (0.9) per sogliare i risultati ottenuti.

Successivamente, è stata implementata una pipeline (Parser) specifica per l'armonizzazione delle variabili di non-imaging coinvolte nel nostro studio, che è attualmente disponibile online sulla piattaforma GitLab. In questo contesto, abbiamo inoltre costruito un modello chiamato Elastic Net e lo abbiamo testato sui dati ricavati dai vari step di ricerca dei parametri ottimali, ottenendo così un valido supporto per il calcolo dell'importanza delle

rispettive variabili di non-imaging. Infine, ci siamo serviti di un regressore gaussiano (Gaussian Process regressor) per la creazione di un modello di predizione generale, in grado di stimare il volume di lesioni cerebrali da cui è affetto un paziente, indipendentemente dalla coorte di dati alla quale appartiene. La performance ottenuta, in termini di correlazione tra il valore attuale e quello predetto, è circa pari a 0.4.

CONCLUSIONI – I dati ottenuti dimostrano l'esistenza di un protocollo generale, in grado di ottenere misure di WMH comparabili tra i diversi dataset a disposizione, nel contesto della segmentazione automatica di lesioni. Tali risultati, insieme al processo di integrazione delle variabili di non-imaging, attestano il raggiungimento di un significativo effetto di armonizzazione sui diversi insiemi di dati coinvolti nella nostra analisi, che risultano finalmente ben integrati e compatibili. La significativa ed evidente eterogeneità che caratterizzava i dataset di partenza consente inoltre di prevedere un'applicazione su vasta scala dell'approccio integrativo da noi sviluppato.

Chapter 1

“Introduction”

In this Chapter we introduce the main aspects of our work, which focuses on harmonisation of imaging-derived measures of White Matter Hyperintensities. We present the clinical context of the condition of interest and provide a general understanding of different harmonisation techniques. Finally, we summarise the main goal we aim to achieve.

1.1 Introduction

The focus of this thesis is the harmonisation (alias, standardisation) of imaging-derived measures of White Matter Hyperintensities (WMHs), which are a common age-related finding on brain magnetic resonance imaging (MRI). WMHs of presumed vascular origin, also known as white matter lesions, white matter disease, or leukoariosis, are areas that appear hyperintense on T2-weighted, fluid attenuated inversion recovery (FLAIR), and proton density-weighted images. Although they do become more common with advancing age, numerous studies indicate that they have important associations with cardiovascular risk factors and clinical impact on cognitive impairment, risk of stroke, risk of functional decline, risk of dementia (Wardlaw et al., 2015; Griffanti et al., 2016).

One of the most important motivations for a further and more detailed characterisation of WMHs is their potential role as early imaging-based markers for cognitive impairment.

With a growing aging population, the burden of people living with dementia represents a major public health issue. Thus, quantitative biomarkers would greatly enhance our ability to detect and ultimately manage dementia, from the early developmental stages, often asymptomatic, up to the actuation of measures of prevention and damage limitation.

During the past few decades several studies have been conducted with this aim (Alzheimer's Disease Neuroimaging Initiative (ADNI), Petersen et al., 2010; The Rotterdam Study, Breteler et al., 1994). However, in order to increase statistical power on one hand and ultimately be able to compare single subject data with general population on the other, we need to integrate datasets, thereby introducing a source of variability.

The distinct properties of different datasets indeed reduce our ability to integrate observations across studies, which would allow us to obtain robust finding to inform developments in prognosis and care.

The main goal of this project is therefore to contribute towards overcoming this barrier by developing approaches to harmonise imaging data from different datasets and different MRI scanners. Particularly we aim at harmonising measures of the volume of white matter hyperintensities in individuals belonging to different populations.

1.2 White Matters Hyperintensities

White Matter Hyperintensities (WMH), defined for the first time in 1985 by Hachinski, are a common sign found on brain magnetic resonance imaging (MRI) or computed tomography (CT) images in elderly subjects or patients with stroke and dementia.

They appear as abnormal areas of signal intensity on magnetic resonance imaging (MRI), in which they are easier to recognise compared to CT, due to the former better sensitivity to soft tissue changes.

Even though the signal change is predominant in the periventricular and deep white matter regions some lesions can also be recognised in the deep gray matter. The presence of these abnormal areas is, however, more significant in the white matter. Many studies show they are bilateral, mostly symmetrical, and hyperintense on T2-weighted (T2), fluid attenuated inversion recovery (FLAIR), and proton density-weighted (PD) images. For these reasons, they are now by consensus referred to as "white matter hyperintensities", where deep gray matter is also involved (Wardlaw et al., 2015).

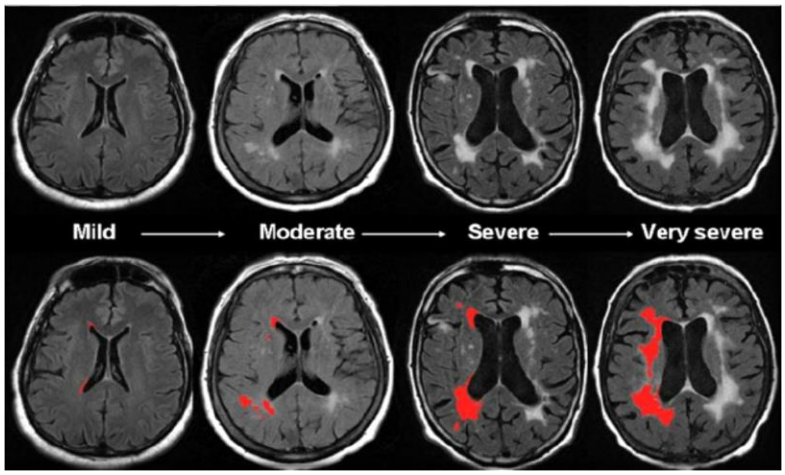


Figure 1.1. Severity of MRI-detected WMH: visual contrast inspection (upper panel) and segmentation (left hemisphere only) by semi-automated tools (lower panel) (Chutinet et al., 2014).

Between the above-mentioned structural sequences, FLAIR is the most sensitive one for detecting WMH but usually they are also commonly found as hyperintense on T2*-weighted sequences and hypointense on T1-weighted ones (Wardlaw et al., 2015). Therefore, all of these imaging techniques are usually involved in the most significant studies relative to WM lesions (Anbeek et al., 2004; Dyrby et al., 2008; Admiraal-Behloul et al., 2005; de Boer et al., 2009).

1.2.1 Pathophysiology

White matter hyperintensities are generally associated to pathological changes in the white matter axonal microstructure or to alterations relative to the interstitial fluid. Examples of these conditions are given by increased water content and mobility, demyelination, or axonal loss that represent permanent structural changes. Extensive WMH were indeed associated with reduced density of glia and vacuolation while more subtle ones with microglial and endothelial activation (Wardlaw et al., 2015).

A further distinction can be made between those lesions mostly occurring around the ventricular structures, called "periventricular" lesions, and those appearing in deep white matter regions, referred to as "deep".

Figure 1.2. Difference between periventricular and deep white matter hyperintensities (Debette et al., 2010).

Periventricular WMH are characterised by discontinuous ependyma, gliosis, loosening of the white matter fibers, and myelin loss around the venules in perivascular spaces. Instead deep WMH are linked to gliosis, demyelination and axonal loss around perivascular spaces, with increased tissue loss as the lesions become more severe.

In most studies, periventricular and deep WMHs are sharply differentiated; however, some others indicate that periventricular and deep WMHs are more probably part of a continuous pathology rather than representing just ischemic and permanent changes due to demyelination and axonal loss (Wardlaw et al., 2015). A recent multi-centre study in 11 stroke centres in China found more microglial activation in WMH in patients with AD than in WMH of age-matched controls, thereby confirming the theory that sees periventricular

and deep lesions as effects of an earlier or later stage in the disease condition affecting patients (Ryu et al. 2014).

After presenting WMHs characteristics and classification, we now mention how they are quantified. There are different approaches to evaluate the overall WMH severity on either CT or MRI, and the most common are the ones listed below:

- Visual rating scales based on the location and severity of the lesions. An example is the Fazekas scale (score 0–3), applied to both periventricular and deep locations (Fazekas et al., 1987).
- Volumetric approach based on manual, semi-automated or automated protocols, using analytical software such as BIANCA (for details relative to BIANCA see Par. 2.4.1, "BIANCA (Brain Intensity AbNormality Classification Algorithm)").

1.2.2 Clinical Context

Although WMHs were once considered just an expected consequence of advancing age, nowadays numerous studies indicate that they have important clinical and risk factor associations with progressive cognitive impairment, risk of late depression, doubled risk of dementia, and a tripled risk of stroke (Wardlaw et al., 2015).

A great amount of research has been conducted so far to try understanding the origin and cause of these lesions, as will be explained in this section.

Kim and colleagues (Kim et al., 2008), showed that presence and severity of WMH have been consistently related to cognitive function and emotion in the elderly. The disruption of subcortical and cortical connections by white matter lesions may indeed be detrimental to cortical gray matter integrity and result in cognitive and emotional dysfunctions.

Debette and colleagues (Debette et al., 2010) demonstrated a significant association of white matter hyperintensities with risk of stroke and age-related diseases. Moreover, several other studies, linked the WMHs presence with subjects cognitive decline (Breteler et al., 1994; Zamboni et al., 2017; Griffanti et al., 2018).

In particular, one of them, the Rotterdam study (Breteler et al., 1994), correlated the presence of WMHs and the enlargement of lateral ventricles with worse performance in all the cognitive tests performed by subjects. Although the presence of white matter lesions and increasing ventricle-to-brain ratio were correlated, their effects on cognitive performance resulted to be independent. After further investigation, the study proposed the conclusion

that white matter lesions on MRI scans represent a morphological substrate of dementia related to vascular disease. However, many of the evidences exploited were based on cross sectional studies and some relevant findings derived from small groups of subjects. Population-based studies including larger cohorts are thus needed to achieve the statistical power necessary to obtain reliable results.

As in this case, several other studies revealed lack of data, not only in the context of clinical and risk factor analysis, but also relevant to the validation of automatic lesion segmentation algorithms, with very small validation samples leading to overfitting (Admiraal-Behloul et al., 2005; Anbeek et al., 2004; de Boer et al., 2009; DeCarli et al., 1995; Dyrby et al., 2008; Ramirez et al., 2011).

To summarise, many neuroimaging studies attempted to analyse the interaction existing between WMH and other pathological conditions thereby trying to achieve a better insight into the impact of WMHs on the subjects state of health. Reaching this goal is of primary importance and could contribute to the prevention of brain damage caused by small vessels disease. Furthermore, it could cast new light over its cognitive and physical consequences relevant to dementia and stroke. However, to achieve this, we must better understand the role of WMHs in aging by integrating observations across studies. This has become possible thanks to the growing number of large cross-sectional cohort studies like Whitehall (Filippini et al., 2014) with 800 subjects, and UK Biobank (http://imaging. ukbiobank.ac.uk), with 100˙000. These large and heterogeneous datasets should be merged together in order to reach the statistical power necessary to provide better, more sensitive and reliable analyses and, crucially, this requires harmonisation.

1.3 Harmonisation

Harmonisation is a process that ensures data compatibility, thus allowing to integrate information from different databases and also to properly explore the similarities and discrepancies across studies. This can thus permit pooling of data from a large number of datasets to obtain statistically valid results.

In terms of harmonisation approaches it’s possible to distinguish between prospective and retrospective ones, depending on when the harmonisation process takes place in the lifecycle of a study. Under prospective harmonisation, the goal is to create an agreement across different studies by determining common measures and protocols before the beginning of

the data collection phase. The reason is that an agreement on a main set of collection procedures and common measures prior to data collection facilitates future integration or comparison of data. Retrospective harmonisation, on the other hand, takes place after the phase of data collection has been initiated, thus giving to all the information gathered in separate studies the label of study specific. It is generally carried out since most of the time datasets are merged at a later stage (i.e. they were not part of the same study at the time of data acquisition), thus impeding to any attempt to prospectively ensure a high level of compatibility across studies, beyond the low-level standards of general application. This retrospective approach therefore necessitates a rigorous documentation relative to all the databases involved in the analysis and a scrupulous a-posteriori process to harmonise and integrate study specific data under a common setup. Despite differences, under both prospective and retrospective harmonisation, the ultimate potential to integrate information is usually related to the same aspects: the level of heterogeneity amongst the different populations involved, the design of the experiments carried out, the characteristics of the machinery involved in the process of data collection and, finally, the standard operating procedures followed by investigators to get all the information needed.

This project mostly aims at performing a retrospective harmonisation process between datasets on both imaging and non-imaging variables, which derive from different machineries and/or were acquired with different tools (e.g. questionnaires or tests) being part of different studies carried out in different centres.

Griffith and colleagues (Griffith et al., 2013) described the general procedure for retrospective harmonisation by three steps. First of all, once a research question guiding the harmonisation initiative has been identified, investigators document all the relevant characteristics of the participating studies, which allows the identification of sources of heterogeneity amongst the available data. This furthermore provides the elements required to properly evaluate the harmonisation potential across studies. Examples of the above-mentioned sources of heterogeneity are usually represented by data access and usage policies, and all the relevant information describing samples, data items, and collection methods (data dictionaries or codebooks, questionnaires, and standard operating procedures). Secondly, based on the documentation obtained and on the scientific purpose of the harmonisation initiative, variables targeted to serve as reference for data harmonisation across studies are selected. A priori selection of the reference variables needs

to balance the trade-off between the desire of integrating as many studies as possible for larger sample sizes and the necessity to limit the heterogeneity of the included studies. Finally, after the identification of the reference variables, there are various methodologies that can be applied to transform all the study-specific data items into the target variable format. These include both qualitative harmonisation and statistical harmonisation methodologies. The former processes study-specific data items using logical algorithms and is often applied to create categorical variables (e.g. alcohol or smoking status). The latter, instead, may be used to harmonise complex constructs, such as cognition.

To conclude, it can be stated there are several methods that allow the integration of different databases but the choice of the best harmonisation approach to use depends on the nature of the measures and overall information to be harmonised. Therefore, it is very important to deeply understand the characteristics of the problem which needs to be addressed and of the available data that would allow to do so (Griffith et al., 2013).

1.3.1 Biomedical Images Harmonisation

Amongst the different types of data involved in our analysis, we had to deal both with non-imaging variables and imaging data, that are particularly sensitive to inter-site and inter-scanner variability. These forms of heterogeneity represent a real problem for joint analyses of MRI data and are usually due to several sources of variability such as the number of used head coils, their sensitivity, the imaging gradient non-linearity, magnetic field homogeneity, differences in the image reconstruction algorithms, as well as many other scanner related factors (Mirzaalian et al., 2016).

In multi-centre studies, a careful prospective harmonisation can mitigate some of the above-mentioned differences thanks to the use of the same acquisition protocol (same pulse sequence parameters and same field strength) and, as far as possible, similar hardware. This allows to work with the same image modalities, all obtained through analogue practical steps.

However, not all the sources of variability can be removed using analogue protocols. Some of them are indeed really challenging, being intrinsic in the image acquisition process when performed using different scanners. Here are mentioned some: the image contrast characterising each scanner, the bias-field affecting MRI images differently from scan to scan, and the scanner dependent Signal to Noise Ratio (Mirzaalian et al., 2016). These

systematic differences, if not entirely removed, can lead to severe biases especially in volumetric analyses and therefore necessitate a retrospective harmonisation approach.
Another source of variability, introduced in the context of automatic lesion segmentation, is given by the analysis method used to obtain the labelling. Distinct raters, heterogeneous training sets as well as the chosen combination of input MRI scans can lead to substantial differences in the volumetric evaluation of the detected lesions.
This is particularly relevant to our project, that deals with WMH volumetric segmentation and therefore will require a correction for the analysis related variability. Furthermore, we address both a multi-centre study, comparing different datasets belonging to the same cohort (Whitehall, see details in Par. 2.6.1, "Whitehall (Whll) phase 11 imaging sub-study"), and data derived from two heterogeneous populations, comparing the Whitehall study to the UK Biobank one (again see details in Par. 2.6.2, "UK Biobank (BB)"). In the first case we will need to correct for the above-mentioned scanner related effects, through a careful retrospective harmonisation approach. In the second one, harmonisation will be even more necessary, since dataset differences will be worsened by lack of a perspective harmonisation process and by the intrinsic heterogeneity of the two populations.

1.4 Objectives

As mentioned above, the main goal of this thesis project is to harmonise measures of the white matter lesional load from heterogeneous datasets (for details relative to all the differences characterising our datasets see Par. 3.1, "Dataset"). In order to pursue this aim, we identified several specific objectives that will be discussed below.
First of all, we aspire to retrospectively harmonise all the non-imaging variables involved in our study (for details relative to our datasets see Par. 3.1, "Dataset"), such as arterial blood pressure, smoking habits, physical activity, cognitive abilities, and many more. This will be done through the creation of a tool able to remove differences caused by heterogeneous data collection protocols that result in variable unit of measurements, different categorisation of the same variables, etc.
Secondly, we aim at harmonising all the imaging variables involved in this work (for details relative to the imaging variables involved in our study see Par. 3.1, "MRI sequences for WMH detection") and we are planning to do so through a dual approach. On one hand, at the pre-processing level, applying specific techniques such as biasfield correction in order

to obtain better comparable contrasts (for details relative to biasfield correction see Par. 2.2.3, “Biasfield correction”). On the other hand, we will address the optimisation of the machine learning approaches implemented in the BIANCA software package, i.e. the automatic segmentation tool used to classify WMH lesions (for details relative to BIANCA see Par. 2.4.1, “BIANCA (Brain Intensity AbNormality Classification Algorithm)”). The objective is to increase its robustness by finding a general setting of analysis parameters able to obtain comparable performances across different datasets. In this way the volumetric amount of lesions relative to different datasets would result similar regardless of the age of the subjects involved in the study. Age is being considered as it is strongly related to the presence of WMHs according to literature (Simoni et al., 2012). Having comparable volumetric distributions in different datasets, throughout the various aging classes will hence be taken as major marker of harmonisation performance.

As part of the imaging harmonisation, BIANCA will be further optimised in terms of accuracy so as to obtain an automatic labelling as close as possible to the manual one. This would allow to widespread the tool usage in many clinical contexts, providing physicians with a robust and reliable support to diagnose dementia, cognitive impairment or any other disease correlated with white matter lesions.

There is one further goal that we aim to reach and that can be identified in the creation of a mathematical model able to predict the volumetric amount of WMHs given all the non-imaging variables mentioned above. This would help us accounting for the variability that is not related to the images themselves but to the demographic or clinical characteristics of the individuals. Furthermore, it would allow us to evaluate the relationship between WMHs and their associated risk factors, thereby providing a better insight into the nature of these lesions.

Chapter 2

"Methods"

As introduced in the previous chapter, WMHs can be found in MRI scans also in healthy elderly subjects, being therefore a diffuse condition and a possible prognostic marker of future neurological and cerebrovascular disorders. Furthermore, the growing number of cross-sectional cohort studies has recently provided a huge amount of data to work with in order to gain a better insight into WMH as risk factor in large population studies. These reasons have enormously increased the interest towards strategies able to improve the WMH automatic segmentation tools born in the past few decades. In this regard, a major issue is to overcome study dependent biases by means of appropriate data harmonisation techniques, thus permitting the integration of different datasets.

This chapter deals with the most theoretical aspects of WMH segmentation and data harmonisation. It also lays the foundation for a deep understanding of our future results, illustrating in detail all the methods involved within the analysis.

Every aspect of the MRI images and of the pre-processing techniques, exploited in the context of our work, will be described in details, along with a proper characterisation of the datasets involved in the study. We furthermore try to introduce the various regressive models and statistical indicators, that will be the basis for the subsequent process of validation of our findings.

2.1 MRI sequences for WMH detection

In this section we are going to introduce all the different MRI images that have been exploited in the context of our work.

2.1.1 T1-weighted images

T1-weighted images are one of the most basic pulse sequences within the MRI field and are due to differences in the T1 relaxation time of tissues. To explain the meaning of this parameter we introduce the following situation: when spins are aligned in an external and constant magnetic field, a radiofrequency (RF) pulse is able to flip them into the transverse plane. Right afterwards, they slide back towards the original equilibrium state, but not all the different tissue types need the same amount of time to do that, each being characterised by a specific relaxation time. Fat, that lies within bone marrow at the brain level, quickly realigns its longitudinal magnetization with the original constant field, therefore appearing bright on a T1 weighted images. Conversely, water and fluids, such as the cerebrospinal one (CSF), have much slower longitudinal magnetization realignment after an RF pulse and therefore appear dark, carrying low signal. Intermediate intensity level are instead produced by white matter and cerebral cortex, appearing respectively light and grey (Westbrook, 2008).

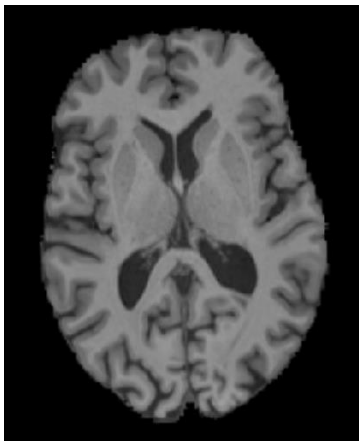


Figure 2.1. T1-weighted image of the brain.

2.1.2 FLAIR (FLuid Attenuation Inversion Recovery) images

Fluid attenuation inversion recovery (FLAIR) is a special inversion recovery (IR) sequence, used in the context of our work as base image for the lesion segmentation tool.

Generally, an inversion recovery sequence begins with a 180° pulse that inverts the net magnetization vector into full saturation. After a specific time, TI, namely the inversion time, an excitation pulse of 90° is applied and transfers magnetization into the transverse plane. The transverse magnetization itself is in turn rephased by an additional 180° pulse to produce an echo. FLAIR utilises a long TI corresponding to the null point of cerebrospinal fluid (CSF) so that the excitation pulse specifically removes the signal deriving from CSF. Brain tissue on FLAIR images therefore appears similar to T2 weighted images, with grey matter (light grey) brighter than white matter (dark grey) and fat that results being bright. Different from a usual T2, CSF, in this case, looks dark.

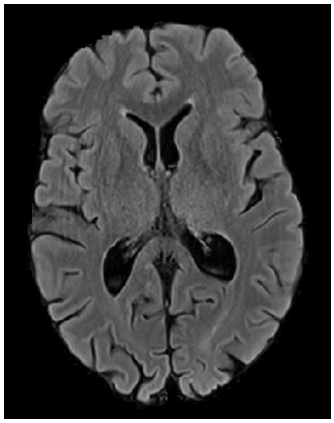


Figure 2.2. FLAIR image of the brain (Westbrook, 2008).

2.1.3 Fractional Anisotropy (FA)

Fractional anisotropy (FA) is a scalar value lying between zero and one that describes the degree of anisotropy characterising a diffusion process. A value of one means that diffusion occurs only along one axis while it is fully restricted along all the other directions. A value of zero, on the other hand, means that diffusion is isotropic, therefore unrestricted in all directions (Alexander et al., 2007).

FA is a measure often used in diffusion tensor imaging (DTI), a particular MRI technique, used to map and characterise the three-dimensional diffusion of water, as a function of spatial location. The diffusion tensor describes magnitude, degree and orientation of the diffusion anisotropy, therefore being able to detect changes in the tissue microstructure and organisation. FA is a parameter particularly sensitive towards these kinds of alterations and is furthermore able to provide information relative to fiber density, axonal diameter,

and myelination in white matter. For these reasons, it was included in the context of our work as one of the available MRI modalities (Alexander et al., 2007).

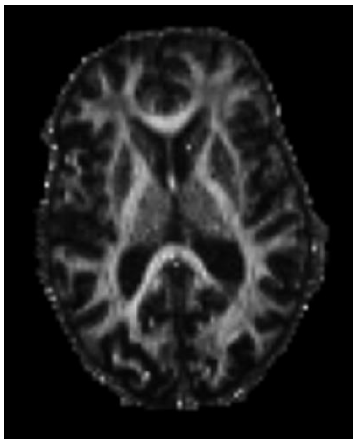


Figure 2.3. FA image of the brain.

2.2 MRI pre-processing

We now introduce some of the concepts that will help pre-processing the MRI scans, preparing them for the further step of WMH detection and segmentation. BIANCA requires indeed some preliminary passages such as brain extraction, registration and bias-field correction before applying the k-NN algorithm to perform the automatic labeling of WMH voxels. Here we briefly describe all of the above-mentioned steps, and eventually the lesion segmentation definition within the BIANCA strategy.

2.2.1 Brain extraction

With image brain extraction we refer to that procedure aimed at deleting all the non-brain tissue from a head volumetric scan. In the context of our work it was performed using a specific tool part of the FSL library (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) called BET (Brain Extraction Tool). This method uses a deformable model that evolves to fit the brain surface by the application of a set of locally adaptive model forces. It is very fast and requires no preregistration or other pre-processing before being applied (Stephen M. Smith, 2002). After deleting all the non-brain tissue, BET can also estimate the inner and outer skull surfaces, and outer scalp surface, when fed with good quality T1 and T2 input images. From a practical point of view, the command takes as input an MRI scan and provides as output the corresponding brain extracted image whose filename needs to be set within the command line. Some additional options allow to obtain further outputs or modifications such as the

following: changes in the fractional intensity threshold from its default value causing the overall segmented brain to become larger or smaller, the generation of a binary mask of the brain, etc.

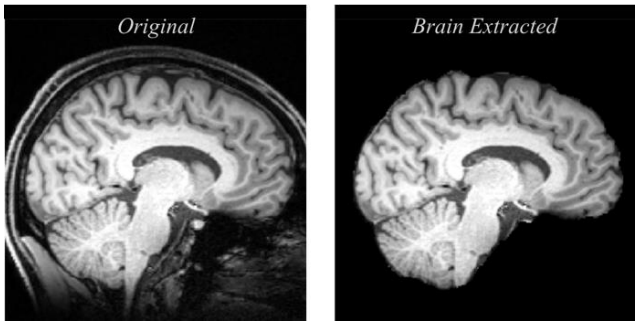


Figure 2.4. Result of the Brain Extraction process (right panel) performed on the original MRI scan (left panel).

2.2.2 Registration

Image registration is a pre-processing technique widely used in the context of medical analysis applications to superimpose images, for further comparisons and/or fusion. The main goal is to reduce them to the same 3D coordinates, so that any other difference or feature overlap will be easier to identify. It is used in motion correction, multi-modal fusion, mapping to Talairach space and many other tasks (Mauren Abreu de Souza et al., 2018).

In most cases images are mapped spatially through the use of automatic registration methods which exploit a cost or similarity function in order to quantify the alignment between two images for any given transformation. The objective function is then optimised in the attempt to find the best solution to the registration problem.

There are several different transformations that can be performed in order to align images and the number of degrees of freedom (DOF) exploited generally characterises the kind of registration implemented. The rigid body transformation is global and linear with 6 DOF in 3D (3 for translation and 3 for rotation). The affine one, is also global and linear and has 12 DOF, adding 3 scaling parameters and 3 shear parameters. The further generalisation to projective transformations (16 DOF) is generally not applied to 3D scans. Non-linear registration permits local (elastic) adaptations of the 3D transformation, by defining a 3D grid of control points, perfectly aligned in the target space, yet on regularly bent lines (e.g., B-splines) in the registered space. Each control point is characterised by 3 DOF, hence the optimisation of a high number of parameters might be necessary, if the grid density was

augmented. For this reason, elastic transformations have to balance the tradeoff between matching of small details and keeping the regularity of convergence, ultimately resulting into image regularity.

Linear transformations are generally used for intra-subject alignment, which means having two (or more) different types of images from the same subject. In the context of our work it was performed using FLIRT (FMRIB's Linear Image Registration Tool), a fully automated tool part of the FSL library (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT). Non-linear registration, on the other hand, is generally used for inter-subject alignment or of subjects to a common atlas. Following a general rule, the best anatomical image (FLAIR) was used to find the affine matching between subjects to MNI atlas. In our case, it was performed exploiting FNIRT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FNIRT) another automatic tool part of the FSL library. Clearly, the intra-subject (subject to atlas) matching of the other modalities is obtained by combining the intra-subject affine transformations with the inter-subject elastic transformation.

2.2.3 Biasfield correction

MR images are often corrupted by a low spatial frequency artifact known as bias field, arising from inhomogeneities in the radiofrequency (RF) field, both in the transmitting and in the receiving phases. Such an artifact causes intensity variations across space, but it has little impact on visual diagnosis, thanks to the ability of the human eye to capture local contrasts rather than the absolute intensity. Unfortunately, this is not equally true in the context of automatic image analysis techniques, whose performance, especially for intensity-based segmentation, can be dramatically degraded by the presence of even mild biasfields (Zhang et al., 2001).

For this reason, within our work, a robust and automatic way of correcting for this artifact is required. In particular, we exploited FAST (FMRIB’s Automated Segmentation Tool), an FSL’s tool able to segment 3D images of the brain into different tissue types (Grey Matter, White Matter, CSF, etc.) and to furthermore correct for spatial intensity variations (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FAST). From a practical point of view the command line takes as input the image to be segmented and returns one binary image for each of the tissue-types classes selected. The different values, from 0 to 1, represent the inclusion probability of each voxel within a particular class of tissue. Further outputs, returned by the

segmentation tool, are a map of the estimated biasfield along with the original but corrected image.

The analytical method exploited by FAST is based on a hidden Markov random field (HMRF) model associated with an Expectation-Maximisation (EM) algorithm. The first one is a stochastic process generated by a Markov random field whose state sequence cannot be observed directly but only through a field of observations. The spatial information contained in each image is encoded through contextual constraints relative to the neighbouring pixels, which are expected to belong to the same class labels. The second, on the other hand, is an algorithm able to derive an estimation of the parameters characterising the original model (Zhang et al., 2001).

Results provided by the HMRF-EM framework are an accurate and robust segmentation of the different tissue types present in the brain, along with the suppression of the bias field due to RF inhomogeneities.

2.2.4 Segmentation

Image segmentation is the process of partitioning an image into different regions based on a given criteria. The ultimate goal is to detect and classify meaningful structures, whether physiological (e.g. cortex vs. white matter) or pathological (e.g., the WMH lesions).

In the context of our work we focused the detection and volumetric quantification of brain lesions in Magnetic Resonance Images, which stresses the central role played by segmentation within this framework. Widely speaking, brain image segmentation is very important for detecting abnormalities such as tumors, edema, necrotic tissues and any other pathological damage. An accurate detection and delineation of these structures is of primary importance for every diagnostic system and is furthermore a key task in many medical applications such as surgical planning, post-surgical assessment, mapping of functional activation onto brain anatomy, analysis of neuroanatomical variability and so on (Balafar et al., 2010).

Brain image segmentation is unfortunately a very challenging task, since brain MRI scans are affected by noise, inhomogeneities, and large inter-individual variability. Namely, the last one goes from the fine details of the individual conformation of gyri and sulci to the gross changes due to pathological sign such as brain atrophy. For this reason, several automatic and general-purpose algorithms have been developed during the past few decades.

For our work, entirely focused on White Matter Hyperintensities (WMHs), we exploited BIANCA (for see Par. 2.4.1), the above-mentioned tool specifically designed for their segmentation.

2.3 Segmentation's previous studies

In this paragraph we will present a review of WMH segmentation approaches that have been adopted in previous studies. WMH segmentation algorithms typically rely on two broad categories of machine learning methods: supervised and unsupervised. Moreover, any algorithm, can be either semi-automated, therefore requiring a certain amount of human intervention at some point during the processing pipeline, or fully automated, therefore requiring no intervention at all (Caligiuri et al., 2015). Several approaches, belonging to all the above-mentioned categories, have been developed during the last forty years. Here we discuss the most important ones.

We start with the fully automated supervised segmentation algorithms, one of the most popular ones being represented by Anbeek and colleagues (2004). Their strategy is to use a k-nearest neighbours (k-NN) classification technique that employs multispectral information from T1-weighted, inversion recovery, PD, T2-weighted and FLAIR scans. Their ground truth is represented by manual segmentations of WMHs. The training set of this study is randomly generated selecting one fifth of the voxels so as to reduce computation time and computer memory. On the other hand, the performance evaluation on a sample is conducted using five different feature sets, each obtained as a combination of voxel intensities from the mentioned MRI sequences, plus the voxel location in the brain. Results show that the best performance is achieved using both intensities and 3D spatial features. Furthermore, the choice of the threshold for the lesion probability map resulting as output has a strong influence on the overall performance of the algorithm. Also, the recruiting strategy of voxels for the training set is shown to have a major impact, since a small number of samples is not representative of whole data (Caligiuri et al., 2015; Anbeek et al., 2004).

Another example of a supervised algorithm, very different from the previous one, is given by the work of Dyrby and colleagues (2008), that faces the segmentation issue by an artificial neural network. Addressed features are: intensities of FLAIR, T1 and T2 images, a 3x3 neighbourhood, and spatial location of each voxel. Furthermore, an optimal-weight-selection strategy is included in the protocol in order to guarantee the classifier generality.

Results from the validation process show that multimodal neural networks outperforms those trained using FLAIR data only and that variations in MRI scan quality represents the largest source of error (Caligiuri et al., 2015; Dyrby et al., 2008).

We now introduce an example of unsupervised segmentation algorithm. Admiraal-Behloul and colleagues (2005) develop a technique comprising two different levels: an adaptive one, robust to differences in image intensities and image contrast, and a reasoning one able to remain unchanged when applied to images acquired on different MRI scanners. Information from three different MRI contrasts were combined in this method: proton density (PD), T2-weighted, and FLAIR. During the first phase, intensity values are mapped to crispy linguistic categories as "bright", "medium-bright", and "dark". During the second one, the linguistic values were used within a fuzzy inference, together with voxel position category (e.g., WM for white matter and IC for Intracranial) to derive a label to every voxel. The innovative aspect that differentiates this approach from others multispectral segmentation methods lies is that different voxel features are exploited only if crucial to the classifier. This results in low dimensionality and thereby in a reduction of the computational time. The method also allows to set some user-defined preferences, such as different exclusion criteria to reduce false positives (Caligiuri et al., 2015; Admiraal-Behloul et al., 2005).

De Boer and colleagues (2009) provide another unsupervised segmentation approach. Their technique allows to segment GM, CSF and WM on multimodal MRI data (T_1- weighted, PD and FLAIR) by an atlas-based k-NN. To train this classifier, a non-linear registration of 12 brain atlases to the subject space is performed. The GM segmentation obtained as output is next used to automatically find a WMH threshold on the histogram of the FLAIR scan. False positives were removed by ensuring that the hyperintensities were within WM (Caligiuri et al., 2015; De Boer et al., 2009).

We now mention the most significant examples of semi-automated segmentation approaches to complete the overview. One is represented by the work of DeCarli and colleagues (1995) which exploits a double-echo pixel intensity histogram to label as WMHs all of those pixels with intensity three or more standard deviations above the mean of the histogram itself. This latter was intensity-corrected according to non-uniformities characterising images (Caligiuri et al., 2015; DeCarli et al., 1995). Another remarkable technique is developed by Ramirez and colleagues (2011) who implement segmentation by applying an adaptive local thresholding. The brain is subdivided in small 3D regions and a threshold is calculated for

each of them, based on the intensity histograms of PD and T_2 images (Caligiuri et al., 2015; Ramirez et al., 2011). In both cases manual steps for checking the quality of the WMHs segmentation is part of the protocol and takes approximately 10–20 minutes of user intervention.

Despite the number of methods proposed, there are several reasons that have prevented them from entering a widespread clinical use, so far. First of all, the code of very few of them is publicly available therefore making very challenging for both researchers and clinicians to evaluate them. Furthermore, several automated and voxel-wise methods have been developed for the detection of multiple sclerosis (MS) lesions, which, however, are significantly different from the WMHs. MS lesions have sharper boundaries and, on the other hand, WMHs are characterised by a very heterogeneous patterns, ranging from punctuate lesions in the deep white matter to large confluent periventricular lesions. A further problem is represented by most algorithms being validated on small samples, with limits related to possible over fitting or to protocol and/or study specificity. These limits severely hinder the overall analysis of completely different datasets (Griffanti et al., 2016).

The need for a multimodal, flexible, freely available and well supported tool led to the development of a new approach. Griffanti and colleagues in 2016 indeed proposed a fully automated, supervised method for WMH detection called BIANCA (Brain Intensity AbNormality Classification Algorithm). The tool relies on a k-NN algorithm, with flexible features (combination of MRI modalities and spatial features) and differs from the previous similar approaches thanks to the introduction of entirely new options like the possibility of weighting spatial coordinates (using local spatial intensity averaging) and changing the number and location of the training points. During this work BIANCA was both optimised and validated in the perspective of the harmonisation of different databases. Firstly, the goal was to find the best combination of parameters (BIANCA options) able to provide a good performance in terms of both overlap and volumetric agreement with a manually segmented WMH mask. Results were evaluated on an annotated subset for each of the datasets involved in the study. The best performance obtained determined the main choices relevant to the settings of BIANCA. In the validation phase, the optimised set of options was applied to the whole cohorts and the resulting measurements of WMH volume were evaluated by correlation with visual ratings and age. Eventually, this study successfully demonstrated that the measure of WMH load extracted with BIANCA is a valid and reliable alternative to

manual segmentation and, furthermore, that the tool could be considered as promising for routine MR diagnostic scans and large cohort cross-sectional studies.
Unfortunately, this doesn't exclude the presence of a certain amount of limitations, that will be further explained in the following paragraph and represent the reason why this thesis project aimed to improve and optimise BIANCA (Griffanti et al., 2016).

2.3.1 Limits

Some of the limitations affecting the above-mentioned approach concern specific steps of the protocol that will be further discussed later in this chapter. Therefore, we just quickly mention them before introducing the most significative one in deeper detail. First of all, there is the necessity to use an exclusion mask of grey matter, cerebellum and subcortical structures to decrease the amount of false positives. Therefore, BIANCA is currently not able to detect cortical and cerebellar abnormalities. Another limit is that the number of clusters k of the k-NN algorithm was not optimised but a value of 40 was selected based on literature. Furthermore, not all the possible configurations of the available BIANCA options were tested, due to an excessively large number of possible combinations.
Last but not least, the training and validation of BIANCA required a training set of manually segmented images, when used with data from different scanners or acquisition protocols. This is the main reason why this thesis project strived to optimise BIANCA both in terms of accuracy and robustness. This happens because FLAIR and T1 characteristics usually vary across different scanners thereby making a study-specific manual labelling mandatory.
The first drawback is that manual segmentation is time consuming and cumbersome, since it requires high expertise in WMH identification. The second consequence, on the other hand, is represented by the lack of a general training set able to obtain comparable performances across different populations. Training BIANCA on one dataset and using it on another one acquired with the same protocol, would further reduce manual intervention and most importantly would make BIANCA applicable to multi-centres studies. That indeed represents one of our main goals (Griffanti et al., 2016).

2.4 Supervised machine learning for segmentation

In this chapter we are going to introduce BIANCA, the automatic tool for lesion segmentation, in greater detail than in the above paragraphs, illustrating its main modules

along with the underlying algorithms. To conclude we will introduce LOCATE, a supervised and automated method sometimes applied to BIANCA's outputs in order to further optimise its performance.

2.4.1 BIANCA (Brain Intensity AbNormality Classification Algorithm)

As previously mentioned, BIANCA is a fully automated and supervised method for WMHs segmentation exploited in the context of our work to perform lesion detection. The tool takes MRI scans of the brain as input, being very flexible about the different image modalities that can be used. It exploits a k-nearest neighbour (k-NN) algorithm in order to classify each image voxel on the basis of intensity and spatial features. The output is further processed to obtain the binary classification mask highlighting lesions (Griffanti et al., 2016).

This is a general overview of BIANCA's working principle. However, in this section we will provide a further insight into the underlying algorithm and its overall structure.

2.4.1.1 k-NN Algorithm

The k-nearest neighbour (k-NN) algorithm is a non-parametric learning method in which it is not needed to discard the training set in the next prediction phase. In fact, the training set has the only function to provide samples that will be compared to validation data in order to help with their classification. Indeed, for each new data to predict, the closest k training samples are selected and the label belonging to the majority of them is assigned to the new example.

In this context, the definition of a similarity measure is of primary importance. This is not always trivial, but it has the advantage of allowing k-NN to be used also with all objects (such as graphs) for which the concept of closest can be defined analytically. A core issue is the choice of the number of classes, defined by the k parameter, which is determinant for the performance of the algorithm and is generally chosen through a cross-validation process.

The k-NN method is affected by the problem of dimensionality, which means that having input data with a very high number of dimensions will decrease the performance of the predictor. This is caused by the fact that, with high dimensions, all of the points tend to have the same distance from one to another (Mitchell, 1997).

When applied to the problem of WMH segmentation, the k-NN method has a feature space whose axis are characterised by voxel features. In BIANCA, these are represented by both punctual, local, and spatial characteristics that can be well summarised as follows: intensity

of the voxel, local average intensity calculated on a 3D patch of pre-specified dimension, and spatial coordinates, respectively.

In order to work, the algorithm requires a training set of pre-classified voxels, each of which corresponds to a specific feature vector. These data, being already segmented, represent a rich set of examples for both the WMH and non-WMH class. Classification of a voxel belonging to a new subject image is then performed through creation of the feature vector, its addition to the feature space, next looking at the k training feature vectors that are closest to it. The output obtained after running k-NN is the probability of each voxel of being WMH and is calculated as the proportion of k neighbours belonging to the WMH class. Its spatial representation is usually referred to as lesion probability map and, being not binary, it needs a further post-processing step in order to produce the final sharp classification. If the probability value exceeds a certain threshold, it is classified as WMH, if not as non-WMH (Griffanti et al., 2016).

2.4.2 LOCATE (Locally Adaptive Threshold Estimation)

To further refine the above thresholding, we now introduce a further automated tool that was used along with BIANCA in the context of our work. LOCATE is a supervised method able to determine adaptive thresholds for binarising the subject-level lesion probability map (LPM). This LPM postprocessing includes the following steps: the division of LPM into sub-regions using Voronoi tessellation, the extraction of local features within those sub-regions, and finally the estimate of optimal local thresholds using a supervised learning method on the basis of the extracted characteristics.

This was a general overview, but we now present in detail the practical steps performed by the tool in order to provide the above-mentioned output. Firstly, local maxima points M_i (i=1, ..., N) are detected on the lesion probability map to identify the plausible lesion locations. Then, the lesion probability map is tessellated, based on those local maxima, into N Voronoi polygons, V_i (for details see Chapter 2, “Voronoi Tessellation”). Within each polygon, different levels of thresholds (Th) from 0 to 0.9 are applied using incremental steps of 0.05, and the following features are extracted at each time:

- Mean greyscale intensity of the image within the thresholded region;
- Distance between the centre of gravity of the thresholded region and brain ventricles;
- Volume of the thresholded region;

The optimal local threshold for each Voronoi region was determined using a random forest regression model (for details see Chapter 2, "Random Forest Regression") with 1000 trees and min leaf size of 5. The training phase was conducted determining, for each V_i, the highest threshold value Th_{max} among the set of thresholds *Th* which gave the best similarity index with respect to the manually segmented binary lesion mask. The random forest regression model was at this point trained with the above-mentioned features using the values of Th_{max} again. In conclusion, during the testing phase the trained regression model was applied to new images in order to get the optimal thresholds *Th* for each of them (Sundaresan et al., 2018).

2.4.2.1 Voronoi Tessellation

Voronoi diagrams were first considered in early 1644 by René Descartes and were further investigated by Voronoi in 1907. They represent a method for partitioning a bidimensional space into convex polygons starting from *n* distinct points. The subdivision is performed in such a way that each polygon contains exactly one generating point p_i and that every point q belonging to a given polygon is closer to its generating point p_i than to any other point in the 2D space. Therefore, the following analytical formula can be used in order to well represent the concept: distance(q, p_i) < distance(q, p_j) for each p_j with $j \neq i$ (Sack et al., 2000). Despite simplicity of the underlying concept, Voronoi diagrams have applications in almost all areas of science and engineering, reaching a widespread use in field such as network analysis, computer graphics, medical diagnostics, astrophysics, hydrology, robotics and computational fluid dynamics.

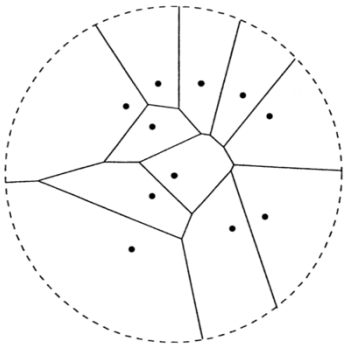


Figure 2.5. A Voronoi diagram of 11 in the Euclidean space (Sack et al., 2000).

2.4.2.2 Random Forest Regression

Random forest, or random decision forest is an ensemble and supervised learning method for classification, regression and other tasks. With the term ensemble, we refer to a technique that combines predictions from multiple machine learning algorithms with the goal of obtaining a more accurate result with respect to the one that would be obtained through an individual model.

Random forest regression is strongly connected with the concept of decision trees and is in charge of correcting for their usual tendency to overfitting.

It relies on a technique called bagging (Bootstrap Aggregation) which is the process of randomly sampling from a dataset with replacement. Therefore, if we have a group of observations of size N, we will be able to generate many new ones, different from the original but characterised by the same amount of data. This technique is very effective when applied to decision trees, being these latter very sensitive to the data they are trained on, for which even small changes can result in significant differences in the structure of the resulting tree.

Random forest regression exploits this concept and operates by constructing a multitude of decision trees during the training phase. Next, each of them processes data during the testing phase and generates a class of predictions, the most voted one being considered as the final RF prediction model (Breiman, 2001).

In conclusion, this algorithm was able to bring significant improvements in classification accuracy, thanks to the many trees protecting each other from their individual errors: while some of them may be wrong, many other will be right, so that, as a whole, they manage to move in the correct direction.

2.5 Harmonisation's previous studies

The following section is split into imaging and non-imaging harmonisation data harmonisation methods used in the past few years with goals similar to ours.

2.5.1 Imaging harmonisation

Even if inter-scanner variability (for details see Par. 1.3.1, "Biomedical images harmonisation") can be minimised acquiring data using the same scanner model and an analogue pulse sequence, recent studies highlighted relevant disparities between

measurement deriving from different sites (Kochunov et al. 2014; Mirzaalian et al. 2015, 2016).

This inter-site variability especially interested fractional anisotropy (FA) images, where it revealed to be uneven being both tissue and region specific (Mirzaalian et al. 2015).

Hence, for joint analysis of data from different sites, three major harmonisation techniques, based on data poling, have been explored during the past few years. The first to be introduced is a *Meta-analysis* approach which relies on the combination of z-scores, relative to distinct sites, in order to establish group disparities. The application example addressed by this study is diffusion tensor imaging (DTI) measures such as fractional anisotropy (FA); nonetheless, harmonisation principles can be extended to other application, WMH studies included. Even though, the subject population at each site may not be adequate to capture the variance of the entire population it is a critical requisite to ensure proper pooling and analysis of z-scores, since the latter depends on the variance and not only on the population mean. A consistent limitation characterising this approach is represented by low statistical power, since z-scores may not be the best method when dealing with non-Gaussian distributions (Bouix et al, 2018).

A second category of methods uses *Statistical covariates* to account for site-specific differences (Forsyth and Cannon 2014; Venkatraman et al. 2015), regressing out the inter-site disparities. The specific applications present attain to FA, again, and to cortical thickness evaluation. The approach introduced by Kochunov et al. (2014) is a mixture of the previous ones and uses z-scores from each site and then regresses-out specific site differences using statistical covariates. Thus, both strategies are integrated to correct for scanner differences through a linear correction factor specific for each of the addressed measures, giving to the harmonisation procedure a model-specific characteristic (Bouix et al, 2018).

Recently ComBat (COMbining BATches), an innovative empirical Bayes approach to correct for batch effect, has been exploited as a retrospective multi-site harmonisation method in many studies about DTI data, cortical thickness measures and other longitudinal ones such as the Adolescent Brain Cognitive Development (ABCD) study (Fortin et al., 2017; Fortin et al., 2018; Nielson e al., 2018).

ComBat works as follows: it obtains a preliminary estimate of the batch interaction parameters from a linear model and then shrinks those parameters towards a grouped mean.

The level of shrinkage is generally calibrated according to either parametric or nonparametric estimates of the batch effects distribution (Johnson et al., 2007).

The main limitation characterising this approach is that specific prior distributions (Gaussian and Inverse-gamma) of the site effect parameters do not allow to properly generalise to every possible scenario or to measures derived from other models.

Moreover, harmonisation of imaging datasets oughts also to permit the exploitability of information recorded aside each image, relevant to the subject anamnestic and demographic features, subject condition at the specific scan, scanner and protocol technical data, etc. All this, by overcoming the different procedures, metrics, and scales adopted in each dataset. The harmonisation approach exploits many different processes, such as: selection of the data to be accepted and further application of a cleaning procedure, application of quality control metrics to the accepted data, mapping of data in order to match them with a common data model, processing of the data with common bioinformatic pipelines, and application of further quality control metrics to the processed results (filtering of poor quality data, outliers removal, etc.) (Lee et al., 2018). These concepts are well summarised in the scheme presented below:

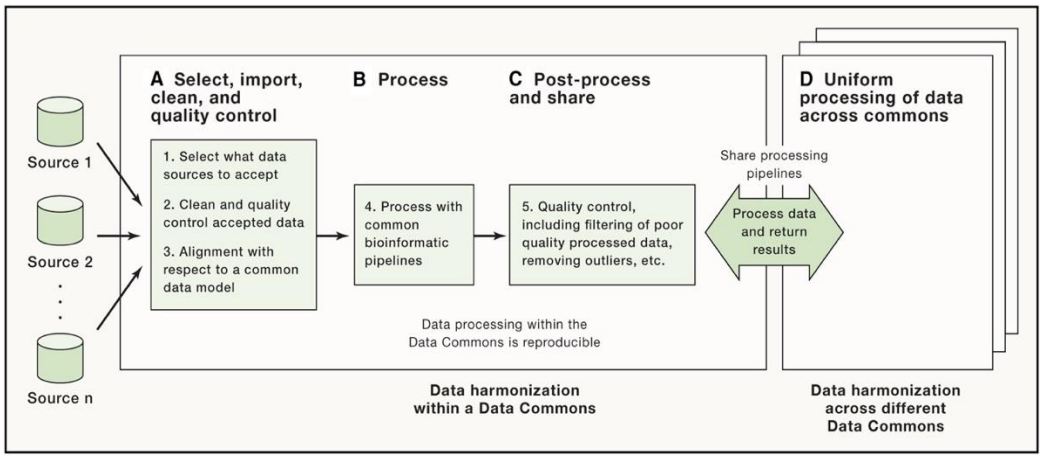


Figure 2.6. An Overview of the Data Harmonisation Process (Lee et al., 2018).

Data harmonisation practices are critical for advancing the knowledge on neurological diseases such as Alzheimer, Parkinson, depression, etc. in a learning health system and, moreover, are important to turn those advances into patient benefit. Literature on statistical methods of data harmonisation mainly contemplates three general approaches. The first one uses a simple linear- or z- transformation to create a common metric, thus combining

constructs measured using different scales across datasets. An example is given by the *Standardisation Methods* that can be applied to continuous variables and do not require a specialised software (Minicuci et al., 2004).

A second class of methods is based on one or more *latent factors,* which underlie a set of measured items that can be modelled through very different techniques: linear factor analysis for continuous items, two parameter logistic item response theory for binary items, a polytomous Rasch model for ordinal ones, and finally moderated nonlinear factor analysis (MNFA) if there is a mix of all of them.

In each of the above-mentioned cases, firstly we need to create a "conversion key" using one of those statistical models to depict the relationship between the latent construct and measured items. The next step consists in converting the information into a common scale through the newly generated conversion key.

On the other hand, the final class of methods, *Missing data by design with multiple imputation,* is applied to continuous, binary and ordinal data but requires some specialised software and multiple datasets (Burns et al., 2011). In this case, the authors were interested in mixing Mini-Mental State Examination (MMSE) scores with missing data across nine Australian longitudinal studies, related to demographic characteristics such as age and education. Burns and colleagues (2011) used a specific model that exploited multiple imputation with chained equations to assign appropriate missing MMSE item scores.

2.5.2 Harmonisation of Non-Imaging data

However, before talking about specific methods for statistical harmonisation, it's fundamental to create inferentially equivalent datasets through a pre-statistical harmonisation process involving the non-imaging data. The latter includes steps such as the identification of relevant cognitive tests, the identification of biological, demographic and clinical variables of interest (e.g., blood pressure, weight, smoker status, sex, age, and education) and finally the qualitative harmonisation of all of the above-mentioned features using processing algorithms able to reduce variables to a common format. Lack of inferentially equivalent datasets would result in qualitative harmonisation being only applicable to simple constructs (e.g., number of cigarettes smoked, sex etc.), but not to more complex measures such as different rating scales across studies (Griffith et al., 2013). This helps understanding the importance of this process.

Inferentially equivalent datasets can be obtained by converting the original data dictionary (a codebook with descriptions of variable names, type, format, and missing values) into a fixed master data dictionary that states overlapping data from all studies either prospectively or retrospectively (Kalter et al., 2019).

An example of this kind of harmonisation approach is represented by *DataSHaPER* (Data Schema and Harmonisation Platform for Epidemiological Research) (Fortier et al., 2011). This method starts with the definition of a set of targeted variables, called DataSchema. A priori rules are then defined for each variable of the DataSchema and are used to establish which data can be validly combined across studies therefore generating a DataSchema variable. Then, the Harmonisation Platform provides a template for formal estimation of the potential to summarise information from different studies. However, this approach is subjects to some limitations that are related to variable selection and pairing rules, participating studies, and the harmonisation process. Moreover, one of the main limits affecting this method is represented by the inevitable element of subjectivity associate to the variable selection and pairing rules definition. Another example of Data Harmonisation Platform (DHP) is the one recently developed for the Predicting OptimaL cAncer RehabIlitation and Supportive care (POLARIS) study, which was subsequently applied to other cohorts (Kalter et al., 2019). This innovative platform allows to harmonise Individual Patient Data (IPD) with a flexible approach, storing data in a centralised and secured database server with large capacity. It allows the user to import distinctive studies data, to harmonise them with a master data dictionary and to further export the harmonised results into a statistical software program of choice (e.g., SPSS statistical package) for further analysis.

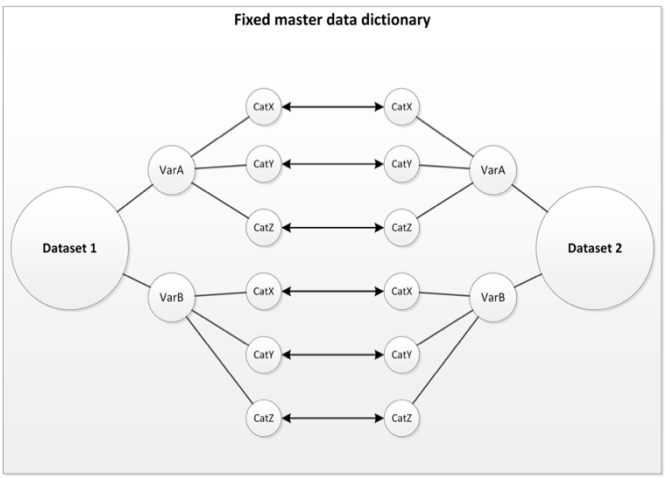


Figure 2.7. Harmonisation model with a fixed master data dictionary (Kalter et al., 2019).

While previous examples dealt with specific fields of interest, working respectively on Epidemiological and Cancer Rehabilitation datasets, this harmonisation work is mainly focused on neurodegenerative diseases. In this specific study, data were taken from the Data portal of the Dementia Platform UK (DPUK) (https://portal.dementiasplatform.uk/). This is a collaboration between DPUK and a growing number of cohort research teams who wish to make their data globally accessible. Indeed, between the DPUK's cohorts we find the ones under study in this project: Whitehall II and UK Biobank, characterised by an inhomogeneous format. For this reason, it was fundamental to process them using an ad-hoc python library called FUNPACK, included in the FSL package.

2.5.3 FUNPACK

As mentioned in the previous chapter, the main disparities between the Whitehall and the UK Biobank datasets are due to differences in the collection of clinical variables, usually happening through different questionnaires or different physical and cognitive tests. There are other differences relative to the type of data collected, being either continuous or integer, binary or categorical. These several disparities made it impossible, so far, to work towards models, comparison and subsequent integration between the two populations with the goal of statistical analysis.

For this reason, it was necessary to introduce a specific pipeline represented by an innovative Python library, designed to convert BB data into a homogeneous format with respect to the corresponding Whitehall ones. This tool is called *FUNPACK* and was developed for pre-processing the UK Biobank data at the Wellcome Centre for Integrative Neuroimaging (WIN), University of Oxford.

FUNPACK comes bundled with metadata about the variables present in BB dataset and contains a large number of built-in rules which have been specifically designed to pre-process those features (https://git.fmrib.ox.ac.uk/fsl/funpack/). The above-mentioned metadata can be obtained from the online BB data showcase (http://biobank.ndph.ox.ac.uk/showcase) whereas the source code is available at https://git.fmrib.ox.ac.uk/fsl/funpack/. FUNPACK aim is to merge one or more input files (e.g. csv, tsv) together and to perform different types of pre-processing, finally producing a single output file. In this thesis work, through the use of FUNPACK, it was possible to

reduce the BB's data into a format compatible with the WhII one. This allowed populations analysis and furthermore the creation of a unique predictive model able to account for the variability of demographic and clinical characteristics relative to the different subjects (for details see Par. 2.7).

2.6 Data management

This thesis project aims to harmonise WMH measurements across two main datasets deriving from the Dementias Platform UK: the UK Biobank (BB) and the Whitehall II imaging cohort (WhII). The cohorts will be integrated through manipulation of imaging data (FLAIR and T1-weighted scans), demographic, clinical and cognitive variables.

The Dementias Platform UK (DPUK), established in 2014, was developed by the Medical Research Council (MRC), in order to improve treatment and ultimately also prevention of dementia (Orton et al., 2018). Especially, it supports multi-modal studies both in genetics and imaging, increasing the chance to ameliorate data collection and data sharing. Concerning the two above-mentioned datasets, both deploy the same MRI contrast types (T1-weighted and FLAIR), same field strength in the acquisition sequences (3T), but distinct resolution and contrast scales. Furthermore, they have overlapping, yet not matching, age ranges (WhII: 65-85 years; BB: 50-80 years) and only part of the cognitive tests submitted to participants in common. In terms of harmonisation, the two different Scanners used within Whitehall (3 T Siemens Magnetom Scanners, SC1: Verio, SC2: Prisma), belonging to a multi-centre study, have been used to derive MRI images using the same acquisition protocol. Therefore, when we compare data belonging to SC1 and SC2, it is possible to assume that at least a partial prospective harmonisation was implemented (Filippini et al., 2014). However, some differences remain between the two datasets due to the distinct post-processing techniques applied to the MRI images; e.g., the bias field correction was performed just on Scanner 2. All of this is no longer the case with the introduction of UK Biobank (using a 3T Siemens Skyra scanner), which is a completely different study. No prospective harmonisation protocol was indeed applied in order to integrate the UK Biobank data with the Whitehall one. Therefore, it was necessary to proceed with a specific pipeline for retrospective harmonisation, as mentioned in the previous chapter.

2.6.1 Whitehall (Whll) phase 11 imaging sub-study

The Whitehall II project is a long-lasting longitudinal study started in in 1985–1988 (phase 1) on a cohort of 10˙308 public workers, men and women, aged 35-55 employed in the London offices of 20 Whitehall departments (Marmont et al., 2004). It provides a remarkable source of longitudinal data to explore factors hypothesised to affect brain health and cognitive aging. In the overall WHII study, participants took part in eleven data collection phases, six of which included a medical screening. Between 2012 and 2013, 6˙035 (age range between 60-85) of the original 10˙308 subjects, participated to the *Phase 11* assessment which included the collection of clinical data that varies from demographic and socioeconomic to biological ones, including cognitive assessment and measurements from MRI scans (Filippini et al., 2014). Table 2.1 well summarises all the different types of non-imaging data that were finally included within the Whitehall cohort.

Demographic	*Age*	*Gender*	*Degree obtained*	
Socioeconomic	*Education*	*Income*	*Work*	*Household composition*
Biological	*Blood pressure*	*Weight, Height*	*Glucose, insulin*	*BMI (Body Mass Index)*
Psychosocial/work exposure	*Effort-reward*	*Demand-control*	*Social support*	*Social networks*
Health behaviours	*Smoking*	*Alcohol*	*Physical activity*	*Diet frequency*
CVD (cardiovascular disease)	*WHO chest pain*	*CVD symptoms*		
General health	*Self-rated health*	*Well-being*	*Medications*	*Quality of life*
Mental health	*General Health Questionnaire*	*Activities of daily living*	*CESD (Centre for Epidemiologic*	

	(anxiety, depression)		*Studies Depression Scale)*	
Health outcomes	*Sickness absence*	*Stroke and myocardial infarction*	*Clinical depression*	*Mortality*

Table 2.1. Example summary of the data collected in the Whitehall II study (Marmont and Brunner, 2004).

Furthermore, from the last phase (*Phase 11)* a sub-cohort consisting of nearly 800 randomly selected WhII participants was extracted: the WhII imaging sub-study. Our project is precisely focused on that dataset, comprising 774 MRI scans, acquired with the same protocol on two 3T Siemens scanners (SC1: Verio, N=551; SC2: Prisma, N=223). The involved subjects were 65-85 years old (Filippini et al.,2014).In conclusion, WhII specifically aimed to explore factors hypothesised to affect brain health and cognitive ageing and therefore has a much more detailed cognitive assessment with respect to the BB, even though they have similar image modalities.

2.6.2 UK Biobank (BB)

The UK Biobank dataset is a prospective cohort study with deep genetic and phenotypic data collected on approximately 500˙000 volunteers from across the United Kingdom, aged between 40 and 69 at recruitment, that happened in 2006 (Sudlow et al., 2015).

After six years, in 2011 a web-based questionnaire was included in the assessment visit towards the end of the recruitment period and two years later a repeated assessment of 20˙000 participants was carried out at the BB Co-ordinating Centre, UK. As final check, most of the Participants were invited to undergo a repeated assessment of all the baseline measures (Bycroft et al., 2018). Hence, this huge dataset aims to improve the prevention, diagnosis and treatment of a wide variety of serious and life-threatening illnesses as cancer, heart disease, diabetes, mental health, dementia. This is pursued through the collection of an extensive range of phenotypic information as well as biological samples. Moreover, answers to the questionnaires were collected during the recruitment, especially on socio-demographic, lifestyle and health-related factors. They completed a range of physical measures, as in the Whitehall dataset, with particular attention to mental disorders such as

bipolarism and schizophrenia. Unfortunately, some of these measures were integrated into the protocol towards the end of the recruitment period thereby not being available for all of the 500˙000 participants. Table 2.2 well summarises all the different types of non-imaging data that were finally included within the UK Biobank cohort.

Demographic	*Age*	*Gender*		
Socioeconomic	*Education*	*Income*	*Work*	*Household composition*
Biological	*Blood pressure*	*Weight, Height*	*Glucose, insulin*	*BMI (Body Mass Index)*
Psychosocial/work exposure	*Effort-reward*	*Demand-control*	*Social support*	*Social networks*
Health behaviours	*Smoking*	*Alcohol*	*Physical activity*	*Diet frequency*
CVD (cardiovascular disease)	*WHO chest pain*	*CVD symptoms*		
General health	*Self-rated health*	*Well-being*	*Medications*	*Quality of life*
Mental health	*General Health Questionnaire (anxiety, depression)*	*Activities of daily living*	*Schizophrenia and Bipolar disorder*	*Alzheimer et al.*
Health outcomes	*Sickness absence*	*Stroke and myocardial infarction*	*Clinical depression*	*Mortality*

Table 2.2. Example summary of the data collected in the UK Biobank study.

With regards to images, it can be stated that an imaging sub-study is actually ongoing on 100˙000 subjects including brain, heart and body MRI, carotid ultrasound, a 12-lead ECG recording and a full-body dual energy X-ray absorptiometry scan. Brain scans are acquired in three dedicated imaging centres, all equipped with identical scanners (3T Siemens Skyra, software VD13) and using the standard Siemens 32-channel receive head coil (Miller et al., 2016).

The huge collection of currently available data provides great statistical power thanks to the high-quality scans of nearly 30˙000 individuals and offers a superb dataset for identifying a normative population distribution. However, the cognitive assessment of BB is limited, preventing a thorough investigation of the associations between WMH and cognitive impairment.

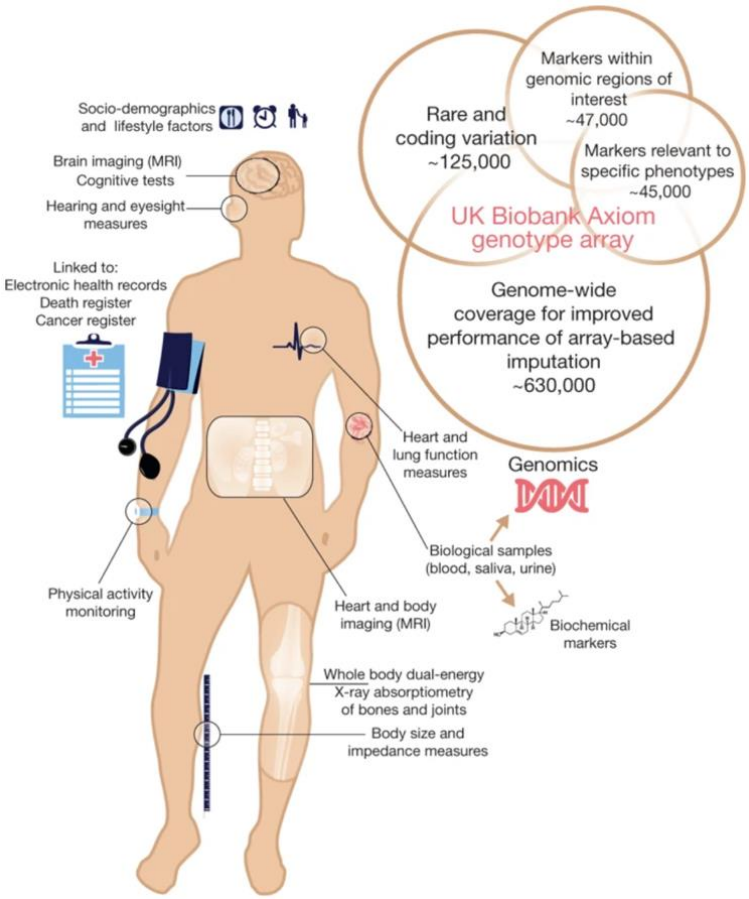


Figure 2.8. Summary of the UK Biobank resource and genotyping array content (Bycroft et al., 2018).

2.7 Predictive models

Once we converted data into a single compatible format, it was finally possible to proceed with the analysis of those datasets. In particular, as previously introduced, we created a unique model able to account for the variability that is not related to the images themselves but to the demographic and clinical characteristics of the individuals belonging to different populations. More in details, the predictive model was generated, firstly, in order to give a general prediction of the amount of WMHs in terms of total volume, starting from just the clinical (included cognitive tests), biological and behavioural data relative to the subjects (e.g. blood pressure, weight, alcohol status, physical activity etc.). This would allow to have a primary idea of the likely presence and severity of lesions, even before resorting to specific magnetic resonance imaging. Secondly, as anticipated, the model has the goal to outline the existing relationship between various non-imaging variables and the total amount of WMH affecting participants, thus highlighting those that have more influence on lesion appearance. These variables could then be indicated as major risks factors. Finally, this would also lead to the observation of similarities and differences amongst the two populations, that basically depend on whether predominant risk factors are the same or not.

A short introduction from predictive analysis and its related models is presented below since it's very common to talk about *Predictive Analysis* in big data, especially in population profiling. This is usually seen as the process of using data analysis to make data-driven predictions, which means exploiting several techniques such as data mining, statistics, modelling and machine learning in order to analyse current data and to build a model for predicting future events (Swani et al., 2017).

Very often, machine learning techniques are used to predict a future value or to estimate probability, capturing relationships among many factors to assess risk with particular set of conditions. Broadly speaking, machine learnings methods can be divided into two sub areas: Supervised and Unsupervised. *Supervised learning* involves allocating labelled data into a system so that a definite pattern or function can be deduced from them. It consists in providing the computer system of the machine with a series of already classified (alias, annotated) examples, that allow to build a database of a-priori information and experience called training set. After the learning phases, which tunes the machine parameters to the best performance on the training set, the classifier will apply the same rules to new data. In other words, it will analyse data based on the previously acquired knowledge (Brownlee, 2016).

On the other hand, *Unsupervised learning* requires that information entering into the machine is not encoded. In other words, the machine has the possibility to draw over the given information without having any reference example, without any knowledge of the expected. Therefore, the machine itself needs to catalogue all the information in its possession, to organise them and learn their meaning, their use and, above all, the result to which they will lead. Namely, the only reference is given by the structure of data themselves; e.g. by the presence of clusters, the separation of which should be optimised. Unsupervised learning allows more freedom in the choice of how to organise information and to learn what is best for different situations (Zhu and Goldberg, 2009).

The main difference between supervised and unsupervised learning is relative to the input dataset. In supervised learning is well known and labelled while in unsupervised learning it is completely unknown (Brownlee, 2016). Another difference lies in the accuracy and in the computational complexity of the results produced after every cycle of machine learning analysis. The results generated from a supervised method are indeed more accurate and reliable when compared to the ones generated from unsupervised algorithms.

In the context of our work, the focus was mainly on supervised learning methods and on a single Bayesian model. Anyway, in Table 2.3 the main algorithms relevant to both will be presented.

	Unsupervised	*Supervised*
Continuous	*Clustering & Dimensionality Reduction* • *SVD* • *PCA* • *K-means*	• *Linear Regression* • *Regularisation (Ridge, Lasso, Elastic Net)* • *Decision Trees* • *Random Forests*
Categorical	*Association Analysis* • *A priori* • *FP-Growth* *Hidden Markow Model*	*Classification* • *k-NN* • *Trees* • *Logistic Regression* • *Naive-Bayes* • *SVM*

Table 2.3. Machine learnings algorithms categories.

2.7.1 General Linear Model (GLM)

As predictive analytics is a tool for machine learning and big data, regression modelling is a technique for predictive analytics. Regression modelling investigates the relationship between a dependent (target) and independent variables (predictor) while also assessing the strength in the association between them. Thus, it is looking for relationships between variables and tries to understand how strong that relationship is. The term general linear model usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors (Goldburd et al., 2016).

In a general linear model:

$$Yi = \alpha + \beta 1X1i + ... \beta pXpi + \mathcal{E}i \quad (2.1)$$

the response y_i ($i=1, ..., n$) is modelled by a linear function of explanatory variables x_j ($j=1, ..., p$) plus an error term. The term general refers to the dependence on potentially more than one explanatory variable, whereas a simple linear model establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line, the regression one:

$$Yi = \alpha + \beta Xi + \mathcal{E}i \quad (2.2)$$

$\boldsymbol{\alpha}$ is intercept, β is slope of the line and $\boldsymbol{\varepsilon}$ is error term. This equation can be used to predict the value of the target variable based on given predictor one(s).

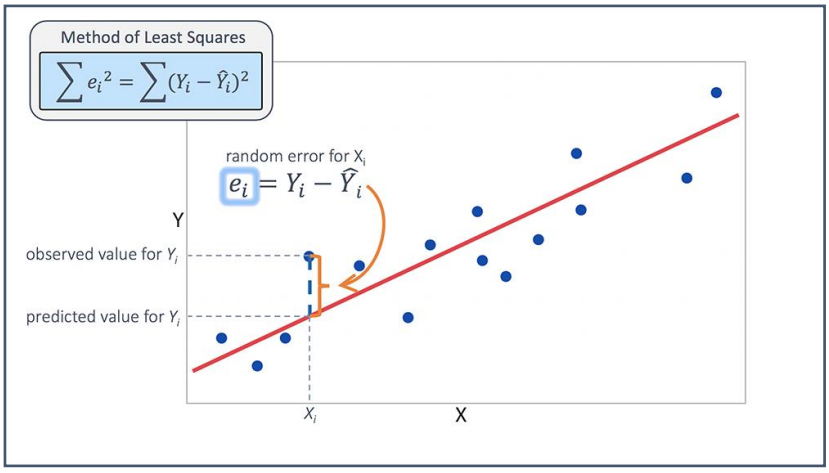


Figure 2.9: Linear regression model, in red regression line (https://www.jmp.com/en_hk/statistics-knowledge-portal/what-is-multiple-regression/fitting-multiple-regression-model.html).

The goal of simple linear regression, also known as Ordinary Least Squares (OLS), is to minimise the sum of squared errors, each error being the difference between the actual data (y_i point) and its predicted value. This quantity, allows to find the model optimal parameters and is generally referred to as cost function:

$$\sum_{i=1}^{n}(Yi- \hat{Y}i)^2 = \sum_{i=1}^{n}(\hat{Y}i - \alpha - \beta * Xi)^2 \tag{2.3}$$

At this point is possible to introduce a further concept. In order to obtain a more accurate model of complex data it's possible to add a penalty term to the cost function since it adds a bias due to model complexity. These penalty terms are generally known as L1 regularisation (Lasso regression) and L2 regularisation (Ridge regression) or a combination of both: Elastic Net.

2.7.2 Ridge Regression Model

Ridge Regression is generally used when independent variables are highly correlated, therefore data suffer from multicollinearity. In this condition, even if the least squares estimate (OLS) is unbiased, the parameter estimate variances are large (H. Duzan et al., 2015). In this regard, it is important to recall that prediction errors can be decomposed into systematic bias and random variance.

Ridge regression is able to reduce the standard error and solve the multicollinearity problem by adding a penalty term to the above-mentioned OLS equation, even though it introduces a certain degree of bias. It exploits L2 regularisation which is the one represented by the following formula:

$$+\lambda \sum_{j=1}^{p}\beta j^2 \tag{2.4}$$

The L2 term is equal to the square of the magnitude of coefficients. When lambda (λ) is equal to zero the equation is basically the OLS model seen before, but when its value increases a constraint is introduced on the coefficients. In particular, this process, called

shrinkage, has the goal of minimising their magnitude, that tends to zero for larger values of lambda (https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net).
In conclusion, shrinking the coefficients leads to a decrease in the overall model variance and in turn results in a lower error value. Ridge regression is therefore able to decrease the complexity of a model without reducing the number of its variables.

2.7.3 Lasso regression

In the sake of completeness, we also describe LASSO (Least absolute shrinkage and selection operator), which is not used in our processing pipeline, yet is one of the most popular algorithms in Machine learning (ML).
It uses an L1 penalty term and stands for Least Absolute Shrinkage and Selection Operator (A. Kassambara, 2018). The penalty is equal to the absolute value of the coefficients magnitude:

$$+\lambda \sum_{j=1}^{p} |\beta j| \tag{2.5}$$

Similarly to Ridge regression, a lambda value equal to zero returns the basic OLS equation, while for higher values a constraint on the coefficients is introduced. The difference is that Lasso regression can drive some coefficients to zero and, in particular, the larger the value of lambda the higher the number of weights that are shrunk to zero. This allows to entirely eliminate some of the features present in the model and provides the selection of a subset of predictors that help mitigate multicollinearity and model complexity. The surviving predictors, whose weights have values different from zero, are hence highlighted as the most important in explaining the variability of the target variable.
In conclusion L1 regularisation is based on the same principle of L2 regularisation but it also allows for feature selection.

2.7.4 Elastic Net

Elastic Net has been developed due to the need of exceeding Lasso regression limits, whose selection of variables can be too dependent on data thus resulting unstable.
The proposed solution is to combine penalties of Ridge and Lasso regularisation in order to get the best of both sites. The overall method aims at minimising the following loss function (A. Kassambara, 2018):

$$\text{Residual Mean Square Error} + \alpha \cdot \text{Ridge Penalty} + (1 - \alpha) \cdot \text{LASSO Penalty} \quad (2.6)$$

In addition to the introduction of a usual lambda variable, Elastic Net also allows us to tune an alpha parameter in such a way that $\alpha = 0$ corresponds to Ridge and $\alpha = 1$ to Lasso. Therefore, it is possible to select a value for alpha between 0 and 1 in order to optimise the elastic Net. Effectively, this will shrink some of the coefficients and set some others to 0 for feature selection (https://medium.com/hackernoon/an-introduction-to-ridge-lasso-and-elastic-net-regression-cca60b4b934f).

We conclude saying that Linear, Ridge and Elastic Net regressions assume a parametric form for functions, differently from Gaussian process models (see nest Par.), which assume a probabilistic prior. The latter brings some benefits, since the uncertainty of estimates is based on statistical inference, but also some challenges, being the algorithms for fitting Gaussian processes more complex than in parametric models.

2.7.5 Gaussian process regression (GP)

GP is a non-parametric (i.e. not limited by a functional form), Bayesian approach towards regression problems, that can be utilised in data exploration and prediction. As a Bayesian technique its approach infers a probability distribution over all possible values useful for regression problems (E. Schulz et al., 2017).

GP generates data located throughout some domain such that any finite subset of the range follows a multivariate Gaussian distribution. Thus, *n* observations in an arbitrary data set $y = \{y_1, ..., y_n\}$ can always be considered as a sample from a multivariate Gaussian distribution.

In contrast with parametric regressions, GP considers every possible function that matches the data, with the drawback of a huge amount of parameters involved. That is indeed the real meaning of non-parametric: it's not the total absence of parameters, but rather the fact that their number is large. To summarise, rather than calculating the probability distribution of parameters for a specific function, GP calculates the probability distribution over all admissible functions that fit the data (https://towardsdatascience.com/quick-start-to-gaussian-process-regression36d838810319) . It operates as follows: first it needs a prior (on the function space), then it calculates the posterior using training data, and finally computes the predictive posterior distribution on the points of interest, thus incorporating information from both the prior distribution and the available dataset.

$$f(x) \sim GP(m(x), k(x, x')) \tag{2.7}$$

The prior selection, or model selection, is a key step, necessary to narrow the range of possible functions to be selected. In this regard the domain of interest is considered, to specify the mean function and the smoothness through the use of a *covariance matrix.* This enables to ensure that values that were close together in the input space will be close in the output space as well.

The mean function is typically constant, being either equal to zero or to the mean of the training dataset. On the other hand, there are several options for the covariance kernel function: it can assume many different forms as long as it follows the properties of kernels. Some of the most common examples include constant, linear, Matern kernel, as well as a composition of multiple kernels.

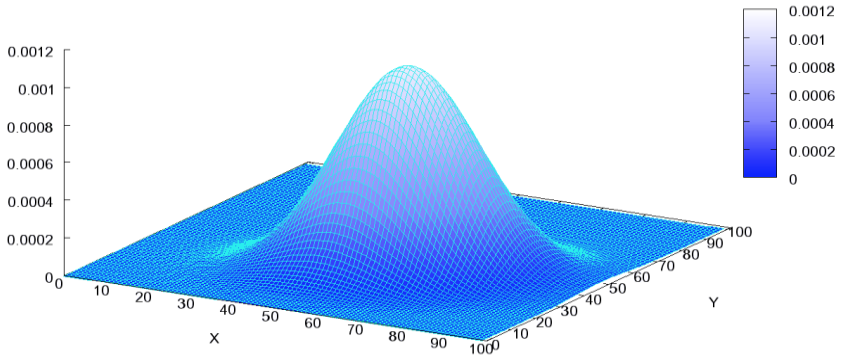


Figure 2.10: Multivariate Normal Distribution (https://katbailey.github.io/post/gaussian-processes-for-dummies/).

In the figure above (Figure 2.10) we can see the bell-shape determined by the covariance matrix (Bailey, 2016). There are several libraries that allow for the implementation of Gaussian process regression (e.g. *scikit-learn*, *GPy*) in Python, but in the context of our work we will focus on the use of *scikit-learn*'s Gaussian process package (https://scikit-learn.org/0.17/modules/gaussian_process.html). A graphical example of the output of GP in *scikit-learn package* is reported in Fig.12. Here observations are represented by blue dots

and a confidence region (95% confidence interval) is plot around the expected predicted function.

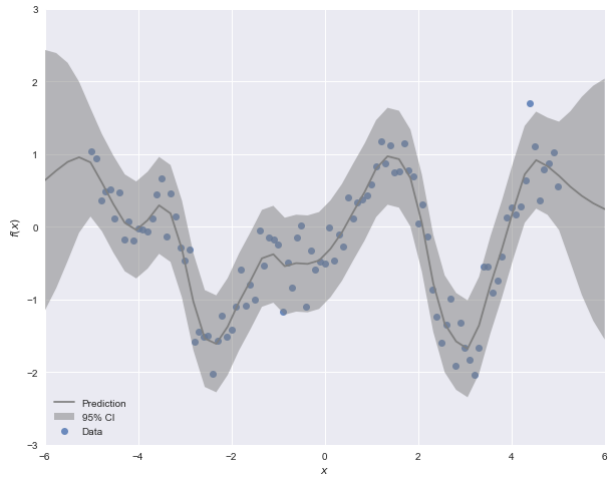


Figure 2.11: GP's graph with confidence interval.

2.8 Model evaluation

Evaluation of model performance is very important. It aims is to quantify the goodness and accuracy of the model on existing data and high performance on future data. Between the model evaluation method, we find the simple analysis of the residuals, which can't give an indication about the quality of new predictions for data it has not already seen, or the best known Validation method: Hold-Out and Cross-Validation.

Evaluating model performance with the data used for training is not acceptable in the context of data science as it can easily generate overfitted models. These last two methods use a test set to evaluate the overall performance. The *Hold Out* method removes a part of the training dataset and uses it to get predictions from the model trained on the remaining samples. More precisely, the original dataset is randomly split into three different subsets:

- A *training set* used to derive the predictive model parameters;
- A *validation set* introduced to evaluate the performance of the model built during the training phase. It provides a test platform for selecting the best-performing model and, moreover, for fine tuning of the model parameters (not all of the available algorithms need it).
- A *test set* used to assess the expected future performance of a model. When an algorithm fits the training set much better than it fits the testing one, overfitting might be the problem (https://www.saedsayad.com/model_evaluation.htm).

Computationally speaking, hold-out validation is simple to program and fast to run. The drawback is its low statistical power when dataset is not large. Unlike the Cross validation, this approach is not able to properly deal with situations in which the available training set has reduced dimensions, therefore introducing the risk of losing characteristic trends in data set and of increasing the error relative to bias.

When a limited amount of data is available, k-fold *Cross-validation* tends to be the best choice to achieve an unbiased estimate of the model performance. For a given hyperparameter setting, the available dataset is split into k folds, each taking turns in being the hold-out validation set. The model is therefore trained on k-1 folds and measured on the remaining held-out one. The overall performance is evaluated by averaging the results given by the k different folds.

A variant to the cross-validation technique is represented by leave-one-out cross-validation. This procedure is based on the same concept discussed above but the value of k is equal to the total number of data points belonging to the original dataset. The testing phase is therefore performed on just one subject (A. Zheng, 2015).

For both K-fold and leave-one-out cross validation, as most of the available samples are used for fitting, the amount of bias is drastically reduced. These methods also lower the value of variance as most of the data are included within the validation set as well. The way in which training and testing sets are swap again and again, also increases the effectiveness of the above-mentioned techniques. As a general rule, values of K equal to 5 or 10 are often used, but this is usually dependent on the specific dataset and computational power available.

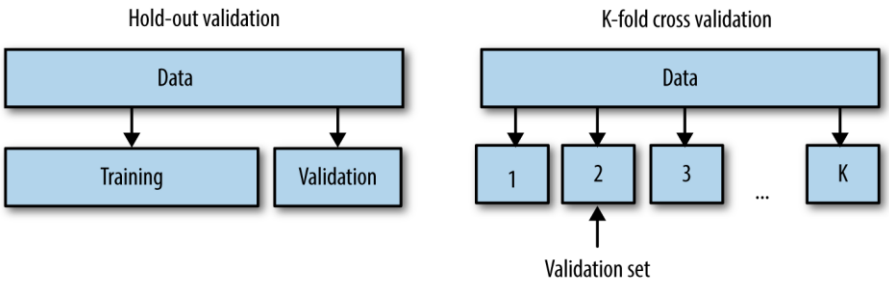


Figure 2.12: Hold-out and k-fold validation flowchart (A. Zheng, 2015.)

2.9 Statistical analysis

After obtaining results from any of the steps involved in our analysis, being able to interpret them is of crucial importance in order to extract knowledge and information that would help

us to deriving significant conclusions about our data. Statistical analysis proved to be very useful in this context and is generally categorised into two main branches: descriptive statistics and inferential statistics.

The first one is defined by a set of techniques used to describe the fundamental characteristics of the collected data and comprises methods such as mean, variance, standard deviation error, maximum, minimum, etc. Together with a graphical representation of data, through the use of scatter plots, box plots, violin plots, and so on, descriptive statistics constitutes the initial basis of any quantitative analysis on data, providing a simple summary of samples distribution and of the collected measures characteristics.

While descriptive statistic is simply presenting what is observed or what data highlight in their essential traits, using information derived from the entire population, inferential statistics, tries to collect one or more samples from the original population and to use them in order to make inferences about the original population. The goal is to derive conclusions that extend beyond the pure data collected process and that may be valid within a wider context (https://socialresearchmethods.net/kb/statinf.php).

Here we present some of the statistical analysis indicators exploited in the context of our work, starting with correlation and T-test, to finally conclude with all the model evaluation metrics.

2.9.1 T-test

A t-test is an inferential statistic method used to establish the existence of a significant difference between the mean of a pair of groups. Hence, it is able to determine whether two sets of data come from the same population or not and, in order to do that, it tests an assumption on the involved populations. For this reason, it belongs to the class of statistical hypothesis tests.

From a practical point of view, the t-test takes a sample from each of the available datasets and assumes, as null hypothesis, that the two means are equal. Based on the applicable formulas, certain values are calculated and compared against the standard ones, leading to either accept or reject the assumed null hypothesis. If it qualifies to be rejected, results indicate the two involved distributions are significantly different from one another. Conversely, if the null hypothesis results in being accepted, it means data actually belong to the same population (https://www.investopedia.com/terms/t/t-test.asp).

There are several types of t-tests, each characterised by a different formula. The most commonly used are the ones that follow:

- *One sample T-test*: used to determine whether a sample of observations could have been generated by a process with a specific mean. Therefore, in this case, we are not trying to compare two distributions between each other, but a single dataset with respect to a specific, fixed value. The corresponding mathematical expression is the one presented below:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{(1+2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.8}$$

- *Independent or unpaired samples T-test*: used to compare the average values relative to two independent groups. In this case we are therefore comparing different distributions between each other.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.9}$$

- *Paired samples T-test*: used to compare samples that represent repeated measures relative to the same group of participants or that are somehow related, having matching characteristics. For these reasons, it can be stated that this kind of test is used to compare directly paired couples of datasets.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2 + s_2^2 - 2\rho s_1 s_2}{n}}} \tag{2.10}$$

In each of the above-mentioned cases, the numerator is represented by the amount of difference between the two mean values (X1 and X2 respectively). Denominator, on the other hand, is obtained through different combinations of samples standard deviation (*s*), samples variance (s^2) and samples size (*n*).

Regarding the *Independent or unpaired samples T-test* a further distinction can be made between *Equal* and *Unequal Variance (Independent) T-test.* The former is used when the involved pair of datasets are characterised by same size or by different size but same variance. The latter, on the other hand, is used for those situations in which datasets are characterised by both different size and variance (Kim, 2015).

2.9.2 Correlation

Correlation indexes are generally used in order to evaluate a possible linear association between two continuous variables. Even though there are several different types of technique that allow to calculate correlation, all indices have certain common characteristics:

1. The values of the various correlation indices vary between -1 and +1; both extreme values represent perfect relationships between variables, while 0 represents a total absence of it. This concept is valid as long as we consider relationships of linear type;
2. A positive relationship means that individuals who obtain high values in a variable tend to get high values even on the second one. In the same way, those who have low values in the first variable tend to have low values on the second one;
3. A negative relationship indicates that at low scores on a variable corresponds a high score on the other variable or vice versa (Mukaka, 2012);

Usually in statistics the most commonly used correlation coefficients are the following:

- *Pearson product-moment correlation coefficient*, also known as *r* of *Pearson*, measures the strength of the linear relationship between normally distributed variables.
- When variables are not normally distributed or the relationship between them is no longer linear, it may be more appropriate to use the *Spearman rank correlation method*, also known as *r of Spearman*.

2.9.3 Evaluation metrics

As discussed in the previous chapter, predictive models can be divided into two broad categories: classification and regression. A classification problem is about predicting the category to whom a training sample belongs to, in the attempt to derive some conclusions from the observed data. According to the category under investigation, different metrics can be used to evaluate the overall model performance. The most common ones are represented by Percent correction classification (PCC), measuring the overall accuracy, and Confusion matrix. The aim of a regression model is to map input samples to continuous real values rather than using classes or discrete variables, as it happened for the classification case. The ultimate goal is the prediction of future values distribution, starting from the original training data. In this context, evaluating model accuracy is an essential part in the process of creating a machine learning model able to perform well. Being $\hat{y}_i$ the predicted value for y_i and $\bar{y}$

representing its mean value, three of the most common metrics used for regression model evaluation are described below (https://indatalabs.com/blog/predictive-models-performanceevaluationimportant?cli_action=1572796194.637):

- *R-squared coefficient*: it represents the proportion of variance characterising the outcome that our model is able to predict on the basis of its own features. This parameter does not take into account the eventual bias that might be present within the data, therefore, a good model is generally associated to a low R-squared value.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{2.11}$$

- *Mean Square Error (MSE*): it represents the average of the squared differences between the predicted outputs and the real ones, useful in case of high number of outliers in the data.

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.12}$$

- *RMSE* (Root Mean Squared Error): is represents the square root of the MSE value.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2.13}$$

Chapter 3

"Materials and Methods"

In this Chapter we'll describe in detail the tools and protocols used in the project. We first explore the effects of the different pre-processing options for obtaining the percent volume of WMH over whole brain (WMH%) in the Whitehall dataset, bias field correction and choice of training data for BIANCA, included. As the Whitehall study was acquired across a scanner upgrade, we also assess the effects of this on our WMH% measures. Then we expand the exploration to the harmonisation of data from both the Whitehall and the UK Biobank, thus broadening the search for optimal parameters able to give comparable results across different scanners. Finally, we will pass through the steps necessary for the construction of a unique predictive model, necessary to validate the improvements related to the performance of BIANCA and to better characterise the relationship between non-imaging variables and WMH% among the different populations under study.

3.1 Datasets

After the introduction of the characteristics related to the general datasets and procedures in Chapter 2, we now present further details relative to the process of subjects selection that was conducted in the context of our analysis.

3.1.1 Whitehall

We start with Whitehall II imaging Sub-study, comprising 774 elderly participants (60-85 years old) acquired with the same protocol on two different scanners. Ethical approval was granted by the University of Oxford Central University / Medical Science Division Interdisciplinary Research Ethics Committee and all participant gave written informed consent.

Scanning was carried out at the Oxford Centre for Functional MRI of the Brain (FMRIB) using two 3T Siemens Magnetom Scanners with a 32-channel receive head coil: a Verio model for 551 participants (Scanner 1: SC1) and a Prisma for the other 223 participants (Scanner 2: SC2).

For each subject, multiple image modalities were recorded, including high-resolution T1 images (MEMPR, TR = 2530 ms, TE= 1.79/3.65/5.51/7.37, flip angle 7°, FOV 256 mm, voxel size 1 mm isotropic), FLAIR images (TR/TE = 9000/73 ms, flip angle 150°, FOV 220 mm, voxel size 0.9 × 0.9 × 3 mm) and fractional Anisotropy (FA). Exclusion criteria complied with the standard MRI safety/quality ones. E.g., metal implants, recent surgery, health conditions problematic for MRI scanning, extreme claustrophobia, inability to travel to Oxford without assistance (Filippini et al, 2014).

A total amount of 34 non-imaging variables were also available for the 774 participants and were subdivided in demographic (e.g. age, gender, weight, height), clinical (e.g. systolic and diastolic blood pressure) and cognitive classes (e.g. Trail Making Test A and B, Digit Coding Test).

Of the 551 subjects recorded on SC1, the following were excluded: 1 due to lack of non-imaging data, 19 because of concurrence with specific pathological conditions such as stroke, cancer and multiple sclerosis, and 3 due to lack of some of the needed MRI modalities. As a result, analysis were performed on a total number of 528 subjects.

Of the 223 subjects recorded on Scanner 2, there were: 4 excluded due to lack of images and 8 because of concurrence with the above-mentioned pathologies. The final number of subjects was therefore reduced to 211.

3.1.2 UK Biobank

The Imaging component of the UK Biobank is an ongoing study, aiming to acquire brain and other MRI scans from 100˙000 predominantly healthy participants, aged 40-69 at baseline recruitment with health outcomes being tracked over the coming decades. Even in this case, informed consent was obtained from all UK Biobank participants and ethical procedures were controlled by a dedicated Ethics and Guidance Council (Miller et al., 2016). Subjects were excluded from scanning according to the same criteria applied on the Whitehall dataset. Of the 14˙503 participants, available at the time of this project, we selected 2˙295 subjects with matching variables to Whitehall. Among those, 10 were next excluded because of concurrence with stroke events. In total, our dataset was composed of 2˙285 patients, all with multiple MRI modalities acquired by a 3T Siemens Skyra with 32-channel receive head coil. including T1 (3D MPRAGE, TI/TR=880/2000 ms, voxel size 1 mm isotropic, sagittal, R=2) and T2 FLAIR (3D SPACE, sagittal, R=2, PF 7/8, fat sat, TI/TR=1800/5000 ms, elliptical, voxel size 1.05x1.0x1.0 mm). Table 3.1 well summarises all the available data presents in this study within the Whitehall and UK Biobank cohorts.

	Subjects	*Manual Masks*
Whll SC1	*528*	*24*
Whll SC2	*211*	*24*
BB	*2285*	*12*

Table 3.1. Summary table of available data.

3.1.3 Manual Masks

BIANCA requires training by manually annotated images. We assessed the effects of using different setting of analysis options on these images, including different raters, and training sets from different datasets.

Concerning Whitehall, a subgroup of 48 subjects, 24 from each scanner, was associated with a manual labelling of the WMH lesions, in the form of a binary mask.
Especially, 24 subjects from the Whitehall cohort imaged using SC1 were manually labelled by two different expert operators, referred to as R1 and R2. Then, a subgroup of 12 of these from each scanner were selected in order to balance the WMH lesional load in the two subgroups.
Further 24 subjects from Whitehall imaged using Scanner 2 were segmented by R2. A third *rater*, R3, performed manual labelling of 12 training images from to the UK Biobank dataset. These concepts are summarised in Table 3.2.

	Manual Masks	*Rater*
Whll SC1	*24*	*R1, R2*
Whll SC2	*24*	*R2*
BB	*12*	*R3*

Table 3.2. Summary table of available manual masks.

The three above-mentioned *raters* were characterised by different levels of expertise being respectively a radiology technician (R1), a medical student (R2) and a neuroimaging researcher (R3).

3.2 Parser

To perform harmonisation, it was essential to combine the variables from the datasets to a standard format. This allowed populations analysis and the creation of a predictive model able to account for the variability associated with demographic and clinical characteristics of the different subjects (for details see Chapter 2, “Predictive models”).
Firstly, we excluded all subjects with non-imaging variables out of the range of the Whitehall population. The intersection of Whitehall and UK Biobank led also to the exclusion of 2 out of the 34 above-mentioned features, due lack of such pieces of information within the UK Biobank dataset.

Once the common variables had been identified, the conversion was conducted through the use of the FUNPACK library. This library has been developed with the UK Biobank dataset. As a result, the library contains a large number of procedures specific to this dataset, allowing it to perform various data sanitisation and processing steps, such as the following (*https://git.fmrib.ox.ac.uk/fsl/funpack/*):

- Columns or rows extraction: selection can be performed through the name of a variable, subjects ID or number of visits (many variables are clinical data repeatedly collected during several different phases e.g. Visit 1.0, 2.0 etc.)
- NA value replacement: values representing indefinite answers can be replaced with NA, (e.g. variables where a value of -1 indicates *Do not know*);
- Categorical recoding: particular categorical columns can be re-coded. For example, variables where a value of 555 represents *half* can be recoded and replaced with 0.5;
- Categorical Binarisation: a column containing categorical labels is replaced with a set of binary columns, each being associated to one of the available categories;
- Child value replacement: NA value can be assigned to those features that depend upon another one in a hierarchical way (e.g. answer to "*How many cigarettes did you smoke yesterday?"* after a negative response to the question "*Do you currently smoke cigarettes?"*);

As FUNPACK provides the ability to define and customise new functions, we decided to define a new pipeline for the Whitehall data. This pipeline is a Parser, that converts between BB and Whll non-imaging variables. It comprises both built-in rules and new functions designed for a specific purpose. This pipeline, developed by us, has now been made available online together with an associated step by step User guide to all users willing to integrate and analyse the WhII and UK Biobank datasets (https://issues.dpuk.org/eugene duff/wmh_harmonisation/tree/master/funpack_wmh_bb). In it, a reader can find:

1. An excel file ("Parser.xlsx") containing all the matching variables between the two datasets with the relative notes and conversion rules;
2. A python file containing all the newly generated conversion functions, not being included in the original FUNPACK library ("conversion.py").

Amongst these rules, we defined functions converting both spatial (e.g. meters to centimetres) and temporal (e.g. conversion from min/day to h/week) units and to perform volumetric normalisations (WMH volume normalised for the total brain volume of the

subject). Due to the different data formats of specific questions in different questionnaires, for some variables it was necessary to create functions able to combine more multiple UK Biobank features in order to create the equivalent Whitehall one. For example, the *Alcohol()* function that allows to convert two categorical UK Biobank variables (unit/day and day/week) into a unique continuous one representing alcohol units per month (Whll format).

3.3 BIANCA

After a detailed description of its underlying algorithm, presented in Par. 2.4.2, we now introduce all steps we followed in order to derive the WMH binary lesion map that represents BIANCA's final output (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Usergui
de). The provided MRI scans underwent specific pre-processing steps (Par. 3.3.1) and the information about the data were gathered into a single text file (Par. 3.3.2) before being fed to the segmentation tool.

BIANCA can be tailored to the specific segmentation purpose through a focused setting of its options (Par. 3.3.3). Finally, since the desired output is usually a binary image, a final step of thresholding is needed to obtain a binary lesion mask (Par. 3.3.4).

3.3.1 Data preparation

As previously mentioned, BIANCA works with multiple MRI modalities, the most common ones being FLAIR (Fluid Attenuated Inversion Recovery) and T1. Sometimes information from diffusion-weighted scans is also included, for example by Fractional Anisotropy (FA) maps. Moreover, the tool is quite flexible and allows either 2D or 3D acquisitions to be included in the dataset.

BIANCA works in single subject space, therefore all the MRI modalities need to be registered to one of them (base modality) and to have equal dimensions (same field of view (FOV) and resolution). The choice of the base MRI modality can be arbitrary, depending on things such as the image quality or the aim of the study being conducted. The first step that needs to be done to prepare data is therefore image registration. In our case, we selected FLAIR as base MRI modality and registration of the other scans (T1 and FA) to its space was performed using FLIRT (Jenkinson et al., 2002; Jenkinson and Smith, 2001) by a rigid transformation (6 degrees of freedom), applying trilinear interpolation.

The second step is brain extraction of at least one of the chosen MRI modalities, to allow BIANCA to derive a binary mask of the subject brain inside which to look for lesions. If we want to further restrict the area were lesions will be detected, thereby reducing the amount of false positives, we can consider using a more restrictive mask. We used make_bianca_mask, a support script released with BIANCA, to create a mask that excludes the cortical GM and the following structures: putamen, globus pallidus, nucleus accumbens, thalamus, brainstem, cerebellum, hippocampus, amygdala. This is because some of these structures can appear hyperintense on FLAIR and therefore risk to be labelled as lesions (cortical and deep GM), while some other are often affected by artifacts on FLAIR scans (e.g. cerebellum).

The affine transformation (FLIRT, 12 degrees of freedom, bilinear interpolation) to normalise the base image to the MNI reference space was also computed. However, the images were left in the subject space, while the normalisation parameters were kept to be delivered to BIANCA.

Finally, BIANCA requires a training set that is represented by a set of pre-classified voxels generally belonging to a manually segmented mask of the WM lesions. As described in section 3.1.1, manual labelling of a subgroup of images was already available for our datasets.

3.3.2 Masterfile preparation

BIANCA needs a text file containing the paths to all the images involved within the analysis. This is generally called *masterfile* and contains a row per subject (either training or query subjects) and, on each row, a list of all addressed files:

- All the images decided to be used for classification (e.g. T1, FLAIR), coregistered to the same base space;
- One brain extracted image that allows to derive a brain mask (or any more restrictive mask);
- The binary manual lesion mask (coregistered to the base space, if needed). For query subjects, that do not possess one, any other "placehold" name can be used in to keep the same column order of training subjects;
- The transformation matrix from subject space to standard space, for subject normalisation. The file of normalisation parameters is usually needed to calculate

spatial features (optional argument), which refer to anatomical position in standard coordinates.

Information order can be arbitrarily modified, as the meaning of each column included in the text file can be specified by BIANCA options (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide).

3.3.3 BIANCA call

BIANCA is executed through a command line comprising several options (full list available at https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide).

In our work we used BIANCA in three use-cases:

- Leave One Out (LOO) testing, which was used on annotated subjects (having a manual WMH mask), included in the training set. This validation procedure is well suited to analyse very limited datasets, as our annotated ones. It consists in the process of excluding 1 subject out of k, in order to use the pool of k-1 for the training and the excluded 1 for the testing, in order to generate an unbiased output that can be used to evaluate the performance of the tool. The procedure is repeated until a result is generated for all of the k involved subjects. An example of command line for LOO is: *bianca –singlefile=masterfile.txt --labelfeaturenum=3 --brainmaskfeaturenum=1 --querysubjectnum=1 --trainingnums=1,2,3,4,5,6,7,8,9, 10,11,12,13,14,15,16,17,18,19,20,21,22,23,24 --featuresubset=1,2 --matfeaturenum=4 --trainingpts=2000 --nonlespts=100000 --selectpts=noborder --spatialweight=2 --patchsizes=3 -o bianca_output.nii.gz -v*. With this command BIANCA will use data from *masterfile.txt*. It will look for information about manually labelled images in the 3rd column of the master file and will derive the binary brain from images in the 1st column. The subject to be segmented is, in this case, the first of the master file (first row) and, since he/she belongs to the group of training participants, BIANCA will use only the remaining 23 to perform the training procedure. The tool will use as spatial features the images in the 1st and 2nd columns and will furthermore extract the spatial features (MNI coordinate) using the transformation matrix listed in the 4th one. Considering the training process, BIANCA's default options were applied, as specified in the following. For each training subject, training voxel subsets were randomly selected with up to 2˙000

voxels within the labelled lesions and up to 10˙000 among the non-lesion voxels, excluding voxels close to the lesion edge (*selectpts=noborder*). A 3D patch with dimension equal to 3^3 voxels was used to perform the local intensity averaging with a spatial weight factor equal to 2. The output image was called *bianca_output.nii.gz*. We used this modality to perform the tests described in sections 3.6.1 and 3.6.2 in order to derive results for those subjects belonging to the training sets.

- Training, performed on the available data eventually saving the training file. Once optimised the parameters on the subjects with manual masks, we saved the generated training file to be applied to the rest of the data. Example call: *bianca --singlefile=masterfile.txt --labelfeaturenum=3 --brainmaskfeaturenum=1 --querysubjectnum=25 --trainingnums=1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18, 19,20,21,22,23,24, --featuresubset=1,2 --matfeaturenum=4 --trainingpts=2000 --nonlespts=10000 --selectpts=noborder --spatialweight=2 --patchsizes=3 --saveclassifierdata=training -v*. The difference with respect to the previous case is represented by the fact that no actual output is generated through this command, therefore the *querysubjectnum* option can refer to any random subject among the ones not included in the training set.
- Testing, performed on unseen data by loading a pre-existing training file. Example call: *bianca –singlefile=masterfile.txt --brainmaskfeaturenum=1 --querysubjectnum=25 --featuresubset=1,2 --matfeaturenum=4 --spatialweight=2 --patchsizes=3 --loadclassifierdata=training -o bianca_output.nii.gz -v*. In this user-case, the options relative to the training phase no longer needed to be specified being indeed substituted by the *loadclassifierdata* argument.
- Training and testing, both performed using the same command. This is a further possibility with respect to ones introduced above, but we didn't use it in the context of our work as, with a high number of subjects, it was computationally less intensive to generate training separately and then apply it to the unseen data.

3.3.4 Thresholding

BIANCA's output is an image mapping the probability of each voxel to be classified as lesions. Being not binary, a thresholding step is needed in order to derive a WMH mask. This can be easily performed using *fslmaths*, a very general image calculator included in the

FSL library. For our analysis we chose a threshold value of 0.9 since it was identified as optimal by Griffanti and colleagues in the context of their analysis (Griffanti et al., 2016).

3.4 LOCATE

As an alternative to applying a global threshold to the lesion probability map, we tested the local threshold adaptation implemented by LOCATE. After the comprehensive description of its working principle, presented in Chapter 2, we now describe the practical steps that need to be followed in order to derive the WMH binary lesion map, representing its final output.

The tool takes as input BIANCA's lesion probability map, along with a few other images (section 3.4.1), in order to extract features that will be used to train the random forest regression. At this point, as it happened for BIANCA, LOCATE can be tailored to specific thresholding purposes through the use of a focused preparation process on the available data. This furthermore implicated the use of different command lines (Par. 3.4.2).

3.4.1 Data preparation

For every subject involved in the analysis the compulsory files are the ones listed below (https://git.fmrib.ox.ac.uk/vaanathi/LOCATE-BIANCA/blob/master/LOCATE_User_Manual_V1.1_20052018.pdf):

- The base image modality used in BIANCA (it is essential to provide at least one image modality), renamed in this way when necessary: <subject_name>_feature_<base_modality_name>.nii.gz. As previously mentioned, in our case it was FLAIR;
- Additional images, that were used as intensity features in BIANCA. They have to be referred to as <subject_name>_feature_<modality_name>.nii.gz when required. In the context of our work, we used just the T1 as an additional feature with respect to FLAIR;
- The binary lesion mask based on manual segmentation and included for training subjects only (<subject_name>_manualmask.nii.gz);
- A ventricle distance map, that is an image whose voxel intensities represent the distance from ventricles within the brain mask (<subject_name>_ventdistmap.nii.gz). This was calculated using the FSL tool

distancemap, of which we present here an example call: distancemap -i <ventricle_mask_image_in_FLAIR_space> -m <brain_mask_in_FLAIR_space> -o <subject_name>_ventdistmap.nii.gz). The <ventricle_mask_image> was already part of our original dataset, but needed to be linearly registered to the FLAIR through the use of FLIRT;

- The lesion probability map (LPM) obtained as output from BIANCA. (<subject_name>_BIANCA_LPM.nii.gz);
- A binary mask of the subject brain (<subject_name>_brainmask.nii.gz). We derived it from the brain extracted modality used in BIANCA;
- A binary mask of white matter, excluding sub-cortical regions (<subject_name>_biancamask.nii.gz;). We used the one introduced in section 3.3.1 and obtained through the use of *make_bianca_mask*;

3.4.2 LOCATE call

As for BIANCA, LOCATE is executed through a command line and can be used for different use-cases (https://git.fmrib.ox.ac.uk/vaanathi/LOCATE-BIANCA/blob/master /LOCATE_ User_Manual_V1.1_20052018.pdf). The process of data preparation can be slightly different according to the specific case that needs to be performed. For Leave One Out (LOO) validation, all of the needed input images have to be gathered within a directory and renamed in the specific, standardised manner highlighted in Par. 3.4.1. The directory itself is then fed to the tool through a command line. In order to perform testing, on the other hand, there is no specific need of moving and renaming files and all of the required images are directly fed into to the tool, providing it with their paths. Here we present details relevant to the above-mentioned user-cases, used in the context of our work:

- Training and testing of the LOCATE model using LOO in order to derive results for all of those subjects belonging to the training sets. An example of LOO call is: *matlab -nojvm -nodisplay -nosplash -r "addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master');addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master/MATLAB');**LOCATE_LOO_testing**('LOO_imgs_directory');"*. This command recalls *LOCATE-BIANCA-master*, a folder containing all of the scripts necessary for the tool to run and, more importantly, the above-mentioned directory obtained gathering and renaming files (*LOO_imgs_directory*);

- Training of the LOCATE model on manually labelled data, saving the resulting file. Example call: *matlab -nojvm -nodisplay -nosplash -r "addpath ('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master');addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master/MATLAB');**LOCATE_training**('Training_imgs_directory');"*. In this case the recalled directory is the one containing data relative to the training subjects;
- Testing of the LOCATE model on unseen data by loading a pre-existing training file. For this user-case we exploited two subject specific commands, combining them with the use of 'for' cycles every time we needed to derive results for a very high number of participants. The first command was necessary to extract a matrix of features based on the involved MRI modalities relative to each subject. Example call: *for i in `cat subject_list-txt`; do matlab -nojvm -nodisplay -nosplash -r "addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master');addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master/MATLAB');**LOCATE_feature_extraction_per_subject**([1,1,1,1],2,'Results','Subject_${i}','FLAIR.nii.gz','T1.nii.gz','bianca_output.nii.gz','biancamask.nii.gz','brainmask.nii.gz','ventdistmap.nii.gz');"; done*. The second, on the other hand, was necessary to derive outcomes on the basis of the feature matrix extracted before. Example call: *for i in `cat subject_list-txt`; do matlab -nojvm -nodisplay -nosplash -r "addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master');addpath('/home/fs0/vbordin/scratch/valentina/LOCATE-BIANCA-master/MATLAB');**LOCATE_testing_per_subject**('bianca_output.nii.gz','/LOCATE_training_files/RF_regression_model_LOCATE.mat','LOCATE_features_${i}.mat',[1,1,1,1],2,'Results');"; done*. The command includes both all of the input files required by LOCATE and the obtained outputs, saved in a folder called *'Results'*.

3.5 Preliminary optimisation of the main training parameters

As previously introduced, BIANCA is characterised by several different options, able to greatly influence its resulting output performance. Since the first part of our work was mainly focused around harmonisation of the two different Whitehall datasets, we tried to assess the best option combination allowing to obtain comparable performances across scanners. The

goal was to make sure that no further bias was introduced by the optional parameters choice, therefore finding a common baseline for the process of data integration.
Among all different options we tested some combinations of the most significative ones, through a step by step process. Default values used in the examples presented in section 3.3.3 are here summarised in Table 3.3.

	Select points	*Non lesion points*	*Training points*	*Patch sizes*	*Spacial weight*
Default	*noborder*	*10˙000*	*2˙000*	*3*	*2*

Table 3.3. BIANCA default options.

We followed the workflow presented below:

- Step1: each value of the *Training points* option was combined with all the possible values of the *Non-lesion points* one;
- Step 2: after fixing the amount of *Training points* to the best result obtain through Step1, we combined each value of the *Non-lesion points* parameter with all the ones relative to *Patch size*;
- Step 3: in the same way we finally tried all the possible combination of *Patch size* and *Spatial weight* parameters, finally selecting the best couple of values obtained;

Training points	*2˙000*	*3˙000*	*4˙000*	*5˙000*
Non-lesion points	*8˙000*	*9˙000*	*10˙000*	
Patch sizes	*3*	*6*	*9*	*3/9*
Spatial Weight	*1*	*2*		

Table 3.4. BIANCA parameters values.

In order to evaluate results, we tested BIANCA's performance on a subsample of manually segmented subjects (12 for SC1 and 12 for SC2) balanced in terms of WMH load. This allowed us to avoid any bias due to different amounts of lesions.

3.6 Evaluation of the influence of different analysis options on WMH harmonisation

In this section, we are going to discuss all of the practical steps that were undertaken in order to evaluate the influence of different analysis options on the output of the segmentation process performed using BIANCA.

For data harmonisation we implemented a step by step process evaluating the following options:

- rater who performed the manual labelling step;
- introduction of the biasfield correction process on FLAIR images;
- composition of the training set used for BIANCA;
- presence of the Fractional Anisotropy as one of the exploited image modalities;
- choice of the thresholding method (global threshold or LOCATE).

The goal was to compare one option at a time, while keeping the others fixed, to understand how a specific parameter impacted on the results.

An across-scanners comparison of the Whitehall data was first focused. Next, the UK Biobank dataset was introduced in the analysis, therefore representing a distinct phase in our study.

3.6.1 Multi-centre study with prospective harmonisation - Whitehall

This first section describes the comparison between *SC1* and *SC2*, as emphasised by the variable *Scanner tested*, presented in Table 3.5. In particular, for every selected combination of parameters, the obtained results were evaluated performing the testing phase on the two subgroups of manually labelled subjects, balanced in terms of WMH load (12 for *SC1* and 12 for *SC2*, as previously mentioned). This allowed us to avoid any bias due to the different amount of lesions characterising the different validation sets.

Table 3.5 presents the different options that were combined in many specific ways within the step by step procedure. The considered training sets were: the 24 manually labelled subjects belonging to SC1 (*TR1*), the 24 relative to SC2 (*TR2*), and a mix of both scanners

comprising respectively all of the subjects (*24+24*) or just the balanced ones (*12+12*). SC1 and SC2, in which training and validation is based on a single scanner, will be referred to as *Study specific*. The two expert raters, used for the annotation, are referred to as *R1* and *R2*. Biasfield correction or its absence are labelled *BC* and *BF*, respectively. *FA* was mentioned when present (*FA*) and substituted with an *X* when excluded and finally, the different thresholding methods that could be used were simply referred to as *Global* (the value was set to 0.9) and *Local* (the valued was decided through the use of LOCATE).

Option tested	*Values tested*	*Scanner tested*
Rater	*R1, R2*	*SC1*
Biasfield	*BF, BC*	*SC1, SC2*
Training set	*Study specific, TR1, TR2, 12+12 (Mixed), 24+24 (Mixed)*	*SC1, SC2*
FA	*FA (with), X (without)*	*SC1, SC2*
Thresholding	*Global (0.9), Local (LOCATE)*	*SC1, SC2*

Table 3.5. Summary table of the available setting of analysis options for Whitehall.

The different steps implemented within the procedure are presented in the following sections.

For each option we evaluated the effect on harmonisation by comparing the results between scanners in terms of BIANCA performance with the metrics described in section 3.7 and in terms of impact of the scanner on the variability in WMH volume using the models described in section 3.8.

3.6.1.1 Rater

We assessed the influence of *Rater*, keeping all of the available options fixed except for one. Therefore, the only difference between settings was represented by the use of manual masks performed, in one case, by Rater 1 (R1) and in the other by Rater 2 (R2). Since masks obtained from different raters were only available for data from SC1 was used for both training and testing. We derived WMH using the following settings:

- TR1 **R1** BF FA Global SC1;
- TR1 **R2** BF FA Global SC1;

Note that for the following tests on the Whitehall dataset we fixed the rater to be R2, since manual masks were available for both scanners.

3.6.1.2 Biasfield

In order to assess the influence of *Biasfield*, for each scanner we generated WMH using FLAIR images not corrected for bias field and compared them against the results obtained when FLAIR was bias field corrected using FAST (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FAST). Therefore, the settings we tested were:

- TR1 R2 **BF** FA Global SC1
- TR1 R2 **BC** FA Global SC1
- TR2 R2 **BF** FA Global SC2
- TR2 R2 **BC** FA Global SC2

3.6.1.3 Training set

At this point we tried to assess the effect that *Training set* had on performance, comparing the five different settings introduced in Table 3.5. In each case the testing phase was performed on both scanners, therefore we obtained the following pairs of data:

- **TR1** R2 BC FA Global SC1
- **TR2** R2 BC FA Global SC2
- **TR1** R2 BC FA Global SC1
- **TR1** R2 BC FA Global SC2
- **TR2** R2 BC FA Global SC1
- **TR2** R2 BC FA Global SC2
- **12+12** R2 BC FA Global SC1
- **12+12** R2 BC FA Global SC2
- **24+24** R2 BC FA Global SC1
- **24+24** R2 BC FA Global SC2

3.6.1.4 FA

We followed an analogue procedure in order to assess the influence of *FA*. This step is justified by the fact that, since FA is not available for all subjects in UK Biobank, the aim is to be able to use only T1 and FLAIR as features for WMH segmentation (as used in the data

currently released). Therefore, as we aim to include BB in our study, it's important to assess the impact that *FA* removal would have on BIANCA's performance.

For this reason, we compared all the settings presented in the previous section with the ones obtained excluding the *FA* from the process of training and testing:

- TR1 R2 BC **X** Global SC1
- TR2 R2 BC **X** Global SC2
- TR1 R2 BC **X** Global SC1
- TR1 R2 BC **X** Global SC2
- TR2 R2 BC **X** Global SC1
- TR2 R2 BC **X** Global SC2
- 12+12 R2 BC **X** Global SC1
- 12+12 R2 BC **X** Global SC2
- 24+24 R2 BC **X** Global SC1
- 24+24 R2 BC **X** Global SC2

3.6.1.5 Thresholding

In order to evaluate the effect that the selected *Thresholding* method would have on results we tried to compare the global one, with probability threshold fixed to 0.9, with the local method, obtained through the use of LOCATE. We therefore applied the latter to all of the settings presented in the previous step and then compared the obtained outputs.

A secondary objective is represented by the possibility to compensate for a possible worsening of performance due to FA exclusion from the set of analysis options. We present here the tested settings:

- TR1 R2 BC X **Local** SC1
- TR2 R2 BC X **Local** SC2
- TR1 R2 BC X **Local** SC1
- TR1 R2 BC X **Local** SC2
- TR2 R2 BC X **Local** SC1
- TR2 R2 BC X **Local** SC2
- 12+12 R2 BC X **Local** SC1
- 12+12 R2 BC X **Local** SC2
- 24+24 R2 BC X **Local** SC1

- 24+24 R2 BC X **Local** SC2

3.6.2 Retrospective harmonisation of Whitehall and UK Biobank datasets

The second part of harmonisation addressed the comparison between the Whitehall dataset, represented by both Scanner 1 (*SC1*) and Scanner 2 (*SC2*), and the UK Biobank, as emphasised by the variable *Scanner tested*, presented in Table 3.6. As for the previous case, in order to avoid any bias due to the different amount of lesions characterising different validation sets, results were validated on the two Whitehall subgroups of balanced subjects, 12 for *SC1* and 12 for *SC2*, and on the 12 manually segmented participants of UK Biobank. The different analysis options, whose influence we are trying to assess in this second part of the analysis, are basically two: the exploited thresholding method, that can be either *Global* or *Local*, and the used *Training set*, adding combinations including UK Biobank (see 3.6.2.1).

Some of the options of Table 3.5 were a-priori fixed, in this second study:

- *FA* was excluded since UK Biobank does not include it;
- A process of *Biasfield* correction was always been applied to the FLAIR images, since the first study had already demonstrated its necessity;
- Due to the lack of a common rater, the manual labelling procedure was performed by different operators for the Whitehall and UK Biobank training data. Therefore, this parameter was set to R2 and R3, respectively.

Option tested	*Values tested*	*Scanner tested*
Training set	*Study, specific, 12+12 (Mixed), 24+24 (Mixed), 12+12+BB (Mixed), 24+24+BB (Mixed)*	*SC1, SC2, BB*
Thresholding	*Global (0.9), Local (LOCATE)*	*SC1, SC2, BB*

Table 3.6. Summary table of the available settings of analysis options for the retrospective harmonisation across Whitehall and UK Biobank.

The different steps implemented are presented in the following sections. Also in this case, we evaluated the effect on harmonisation by comparing the results in terms of BIANCA

performance with the metrics described in section 3.7 and in terms of impact of the scanner on the variability in WMH volume using the models described in section 3.8.

3.6.2.1 Training set

With this step we tried to highlight the effect that mixed *Training sets* would have on performance and we compared their results with the ones obtained from the *Study specific* case. We added to the previous options (Study-specific, *12+12* and *24+24)*, two further training sets, consisting in a combination of the Whitehall dataset with the UK Biobank one: *12+12+BB* and *24+24+BB*.

3.6.2.2 Thresholding

In an analogue way to has been done in the first part of the analysis, in this step we are trying to assess the effect that the selected *Thresholding* method would have on results. In particular, we want to understand whether the application of LOCATE on the available data could result in an improvement of the segmentation performance. For this reason, we compare all of the settings presented in the previous step with the ones obtained through the application of a local thresholding method.

3.7 Indicators for the evaluation of WMH segmentation

In order to evaluate WMH segmentation results, we used performance indicators of the overlap between the automatic WMH segmentation and the manually labelled WMH mask:

- *Dice Similarity Index* (DI): calculated as twice the amount of voxels lying within the intersection of automatic and manual masks, divided by the sum of manual mask lesion voxels (true WMH voxels) and tool lesion voxels (positive WMH voxels);
- *Voxel-level false positive ratio* (*FPR*): number of voxels incorrectly labelled as WMH (false positive, FP) divided by the total number of voxels labelled as WMH by the tool (positive WMH voxels);
- *Voxel-level false negative ratio* (*FNR*): number of voxels incorrectly labelled as non-WMH (false negative, FN) divided by the total number of voxels labelled as WMH in the manual mask (true WMH voxels);
- *Cluster-level FPR*: number of clusters incorrectly labelled as WMH (False Positive clusters) divided by the total number of clusters labelled as WMH by the tool (positive WMH clusters);

- *Cluster-level FNR*: number of clusters incorrectly labelled as non-WMH (False Negative clusters) divided by the total number of clusters labelled as WMH in the manual mask (true WMH clusters).

All these measures of overlap, were calculated in the reference space, represented by FLAIR image modality. Since the Dice Similarity Index is the most widely used metric, able to account for repeatability in validating medical volume segmentations, and is given highest importance for the final discussion. Also, cluster level FNR was privileged compared to FPR, since it is known (Griffanti et al., 2016) that cluster level sensitivity is more important than specificity for lesion detection.

After evaluating the segmentation performance using the indicators discussed so far, we tried to compare results between each other. The goal was to identify the best setting among the ones tested within each step. Sometimes we also tried to compare results obtained applying the same setting on different groups of data (SC1 vs SC2, SC1 vs SC2 vs BB, and so on). This allowed us to understand whether the selected combination of parameters could help harmonising data across datasets, therefore providing uniform and integrated results.

3.8 Predictive model construction

In this section we go through all of the steps performed to build the General Linear model and the other model as the regularisation ones (Ridge Regression and Elastic Net), described in detail in the previous chapter "predictive models".

Our aim was to build regression models that predict the percentage lesional load (WMH%) from a individuals non-imaging variables. In addition, these models should help to reveal the inner relationships occurring between specific subgroups of variables and WHM%.

This work was divided in two parts:

1. The first one deals with making these predictions in the Whitehall dataset, since it is characterised by two distinct Scanners (*SC1* and *SC2*), but with same population and MRI acquisition protocols;
2. The second combines Whitehall and UK Biobank, therefore dealing with completely different datasets characterised by distinct populations and different MRI protocols;

Its construction required the use of *Pandas*, a fundamental high-level building block for doing practical, real world data analysis in Python (https://github.com/pandas-dev/pandas).

3.8.1 Multi-centre study with prospective harmonisation - Whitehall

The search for an optimal model began with the implementation of a basic *General Linear Model* (GLM) using data relative to the Whitehall Cohort (*SC1* and SC2). We applied this to the output of BIANCA with the following training parameters: single training set (Training 1), Rater 2, no biasfield correction, FA still present and global thresholding method (*TR1 R2 BF FA Global Thresholding*).

3.8.1.1 GLM

The GLM was implemented assigning the designated 34 non-imaging variables to the independent input variables (X_i with i=1,2,3…34) and the percentage brain lesional load (WMH%) to the dependent output one.

Initially, we fitted the model using both *SC1* and *SC2* data, split into Training (75%) and Test set (25%). Data shuffling was applied to reduce variance and ensure models remaining as general as possible, preventing overfitting.

3.8.1.2 Other Models

We also explored:

1. Decision tree Models, which are better at capturing the non-linearity in the data:
    - Random Forest: an ensemble of different regression trees used for nonlinear multiple regression, where each leaf contains a distribution for the continuous output variable/s.
2. Regularisation Models, which prevent overfitting by extending the cost function to include the goal of model simplicity. They address some of the drawback characterising GLMs by imposing a penalty on the size of the prediction coefficients:
    - Ridge regression: it uses L2 regularisation to decrease the coefficients value but is unable to force them to zero. This severely limits the use of this regularisation technique as it's unable to perform feature selection;
    - Elastic Net regression: it includes both L1 and L2 norm regularisation terms, potentially representing a model that is both simple and capable of performing features selection/reduction.

Finally, once the optimal prediction model was identified, it was used to assess the impact that different training parameters had on performance. In particular, we quantified the effect of bias field correction and different thresholding methods (both *Global* and *Local*) on model

performance. We also selected a subgroup of predictive features to be later used in our analyses.

3.8.1.3 Evaluation of harmonisation through predictive modelling

We also explored the prediction of the Fazekas total score (the Fazekas score is a qualitative score from 0-6 used by expert raters based on visual assessment of periventricular and deep lesional load), to determine how this varied from the output of BIANCA. We again used the first available set of processed data (*TR1 R2 BF FA Global Thresholding*) and the optimal predictive model. We were interested in the most predictive features obtained in both cases and were interested in determining whether BIANCA's output was more independent of the scanner used (*SC1*, *SC2*) than a non-automated visual rating scale such as Fazekas.

After fitting these models, we began a process of evaluation of different BIANCA settings of analysis options (see Table 3.7). Specifically, we tested:

1. *"Single Training (TR1), Rater 2, FA, BF, Global Threshold (0.9)"* versus *"Single Training (TR1), Rater 2, FA, BC, Global Threshold (0.9)"* to evaluate the impact of biasfield correction;
2. *"Single Training (TR1), Rater 2, FA, BC, Global Threshold (0.9)"* versus *"Mixed Training (24+24), Rater 2, FA, BC, Global Threshold (0.9)"* to evaluate the impact of a mixed training;
3. *"Mixed Training (24+24+BB), Rater 2, FA, BC, Global Threshold (0.9)"* versus *"Mixed Training (24+24+BB), Rater 2, FA, BC, Local Threshold (LOCATE)"* to evaluate the impact of different thresholding methods.

The obtained results were graphically represented through the use of different scatter plots, in which data were differentiated according to the variable *Scanner of test*. This allowed us to visualise if and how each combination of parameters helped harmonising data: a positive effect resulted in more uniform and integrated distributions, while a negative one was represented by a further distinction between scanners.

Training set	*Rater*	*Biasfield*	*FA*	*Threshold*	*Scanner of test*
TR1	*R2*	*BF*	*FA*	*Global (0.9)*	*SC1, SC2*
TR1	*R2*	*BC*	*FA*	*Global (0.9)*	*SC1, SC2*

24+24 (Mixed)	*R2*	*BC*	*FA*	*Global (0.9)*	*SC1, SC2*
24+24+BB (Mixed)	*R2*	*BC*	*FA*	*Global (0.9)*	*SC1, SC2*
24+24+BB (Mixed)	*R2*	*BC*	*FA*	*Local (LOCATE)*	*SC1, SC2*

Table 3.7. Summary table of all the settings of analysis options exploited for the evaluation of harmonisation through predictive modelling.

3.8.2 Retrospective harmonisation of Whitehall and UK Biobank datasets

The second stage of modelling explored prediction of WMH% across Whitehall and UK Biobank. With the introduction of the UK Biobank dataset, the input variables reduced from 34 to 32 features, since the number of medications for depression and the years of education missed in many cases of the UK Biobank.

3.8.2.1 Gaussian Process Regression

We finally introduced the Gaussian process (GP) regression, a supervised machine learning approach that can model non-linear relationships between the non-imaging variables and WMH volumes. The GP returns the prediction of WMH in terms of both mean and variance. This helped us to:

- Build a common model able to predict the confidence interval of brain percentage lesional load (WMH%) for a new patient (new testing sample) based on non-imaging variables.
- Compare the different WMH volume distributions segmenting by the most predictive variables.

The complexity of the non-parametric GP model grows together with the size of the dataset. For instance, when applying a Gaussian process to a dataset of size N, exact inference has computational complexity $O(N^2)$ with storage demands of $O(N^3)$ (Hensmanet al., 2013). Therefore, a subset of input variables was selected, in order to avoid predictive power loss and improve model generality. The subset was common to Whitehall and UK Biobank and derived from the intersection of the most predictive variables identified during previous step. In conclusion, given the greater statistical power provided by UK Biobank with respect to Whitehall, we decided to assess whether the GP model, trained on UK Biobank and tested on Whitehall, could give comparable result to the ones obtained both training and testing on

Whitehall. A positive result would provide evidence of the prediction model robustness vs. the introduction of further datasets.

3.9 Statistical analysis

Finally, for every model implemented we introduced a K fold cross validation to overcome the problem of scarce generalisation capabilities. In the context of our work we used K=5. Typical values are either k = 5 or k = 10, as they have been empirically shown to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance (James et al., 2013).

Then to assess the accuracy of the obtained model, some statistical indicators (more details in Chapter 2 "Indicators for statistical analysis") have been calculated. In particular:

- R-squared coefficient, which accounts for the proportion of variance explained by the predictors;
- Root-Mean-Square-Error (RMSE) which provides the magnitude of the prediction errors as a single measure of predictive power;
- Spearman correlation coefficient between the predicted and the actual value, which account for the statistical dependence between the rankings relative to those two values;

Actually, in the context of our work we mainly focused on *Spearman Correlation.*

As previously mentioned, it is a non-parametric index that has a rather simple calculation method: it operates a transformation in which values are replaced by their rank, when data are sorted. Its calculation and subsequent significance testing require the following assumptions to hold (Myers et al., 2006):

1. All of the involved variables need to be either interval or ratio level or ordinal;
2. All of involved variables need to be monotonically related;

Here we present its formula, where n is the number of observations involved and D the difference between ranks:

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)} \tag{3.1}$$

Moreover, in the context of evaluation of different analysis options influence, we assessed the significance of comparisons of outcomes in the validation sets, we applied t-tests.

Chapter 4

"Results"

In this Chapter we go through the preliminary workflow developed to optimise the main parameters of the automatic segmentation tool involved in our work (BIANCA.) Furthermore, we present results of the main analysis steps that allowed us to integrate and harmonise the different datasets involved in our work.

4.1 Preliminary optimisation of the main training parameters

Ahead of the detailed assessment of each analysis parameter effect, described in Par. 4.4 and relevant to a mix of the Whitehall and the UK Biobank datasets, here we summarise the main settings of BIANCA training. This preliminary analysis was limited to Scanner 1 (SC1) and Scanner 2 (SC2) of the Whitehall II data only and was necessary in order to find the best set of BIANCA options able to provide comparable performances across scanners. Since the Parser, needed to integrate the UK Biobank set, was not necessary in this phase, this procedure is presented below in Par. 4.2. Furthermore, the present optimisation is based only on comparison statistics (Dice Similarity Index, False Positive Ration, False Negative Ration, Cluster-level FNR and Cluster-level FPR), without the prediction model based on non-imaging data, illustrated in Par. 4.3.

The analysis of BIANCA training parameters for all possible combinations was performed through the following steps:

- Step1: the value for the *Training points* option that gave the best result (for all the possible combinations of *Non-lesion points)* is 2˙000;
- Step 2: the value for the *Non-lesion points* option that gave the best result (for all the possible combinations of *Patch size)* is 10˙000;
- Step 3: in an analogue way we finally tried all the possible combinations of the last two parameters, being *Patch sizes* and *Spatial Weight*, obtaining respectively optimal values equal to 3 and 2;

In the sake of briefness, detailed results relevant to all the tested combinations are omitted. So, in Table 4.1 we summarise only the values of DI and Cluster-level FPR. Also, each row of Tab. 4.1 presents the outcomes for tested values of a single parameter, when the other ones are fixed to their optimum. Given the purposes described above, results relative to both SC1 and SC2 are reported for all cases and compared.

In this context, for every step involved in the analysis, we identified as optimal the BIANCA training option characterised by statistical performance indicators able not only to provide high values for both scanners, but also to minimise the difference between them. The values, presented in Tab. 4.1 and highlighted in blue, therefore represent the best compromise found.

Step 1	***Training points***	***2˙000***	***3˙000***	***4˙000***	***5˙000***
Scanner 1	*Dice Index*	*0.672*	*0.656*	*0.651*	*0.643*
	False Positive Cluster	*0.632*	*0.709*	*0.755*	*0.774*
Scanner 2	*Dice Index*	*0.7430*	*0.7534*	*0.7504*	*0.7417*
	False Positive Cluster	*0.4385*	*0.4950*	*0.5896*	*0.6422*
Step 2	***Non-lesion points***	***8˙000***	***9˙000***	***10˙000***	
Scanner 1	*Dice Index*	*0.659*	*0.665*	*0.672*	
	False Positive Cluster	*0.663*	*0.656*	*0.632*	
Scanner 2	*Dice Index*	*0.756*	*0.750*	*0.743*	
	False Positive Cluster	*0.486*	*0.452*	*0.439*	
Step 3A	***Patch sizes***	***3***	***6***	***9***	***3/9***
Scanner 1	*Dice Index*	*0.672*	*0.675*	*0.680*	*0.684*
	False Positive Cluster	*0.632*	*0.398*	*0.436*	*0.377*
Scanner 2	*Dice Index*	*0.743*	*0.751*	*0.752*	*0.761*
	False Positive Cluster	*0.439*	*0.369*	*0.302*	*0.356*

Step 3B	***Spatial Weight***	***1***	***2***		
Scanner 1	*Dice Index*	*0.664*	*0.672*		
	False Positive Cluster	*0.741*	*0.632*		
Scanner 2	*Dice Index*	*0.510*	*0.743*		
	False Positive Cluster	*0.154*	*0.439*		

Table 4.1. Summary table of the main settings of BIANCA training. Each value is associated with performance indicators (Dice Similarity Index and Cluster-level FPR) for both scanners. Optimal values highlighted in blue.

In summary, the optimal set of parameters was the same for SC1 and SC2; i.e., 2˙000 *Training points* = 2˙000, *Non-lesion points* = 10˙000, Patch size = 3 and Spatial weight = 2, which gave maximum DI = 0.672 (median) for SC1 and DI = 0.743 for SC2. The Cluster FPR was 0.632 for SC1 and 0.429 for SC2. The fairly high % of voxels in false clusters is justified by our need to privilege sensitivity over specificity in setting the global threshold of positive classification probability, which ought to be not too high in order not to miss parts of the true lesion. However, this figure of loss can be easily improved by a postprocessing excluding small isolated clusters and also by the application of a locally modulated threshold (LOCATE).

Two examples of visual assessment of segmentation are presented in Fig. 4.1. In the top row (Fig.4.1 a-d) a segmentation with non-optimal setting is presented, while in the bottom one (e-h) a good fitting with optimal parameters is shown. From left to right, Fig.4.1 displays the FLAIR image (a and e), the image with BIANCA's segmentation (red, b and f), with the manual mask (white, c and g), with both masks superimposed (d and h).

The top row segmentation shows that the non-optimal setting gave a scarce lesion segmentation performance, as the overlapping region between the masks highlighted, respectively, in red and white is very narrow. On the other hand, the bottom row displays

results of a good performance, as BIANCA output matches almost entirely the manually segmented region.

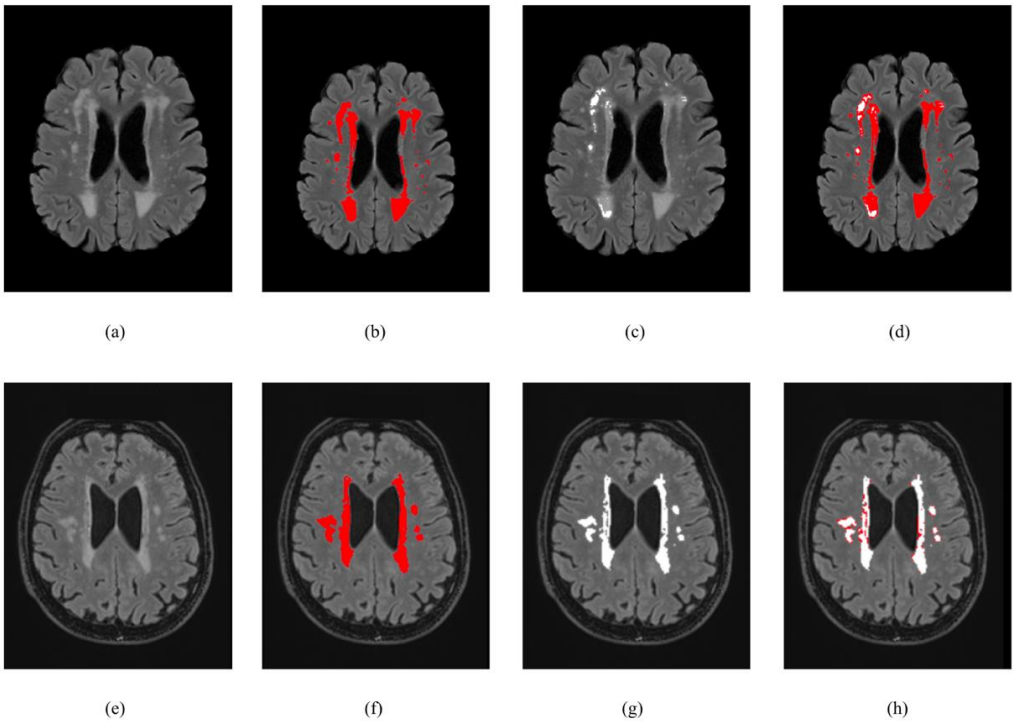


Figure 4.1. Comparison between manual and automatic segmentation performances. FLAIR image characterised by a high (a) and moderate (b) lesional load, relative manual segmentations (b, f), automatic lesion segmentations obtained through the use of BIANCA (c, g) and overlap between manual and automatic lesion masks. This first (a, b, c, d) and second set of images (e, f, g, h) give, respectively, an example of scarce and good BIANCA performance.

4.2 Parser

In this section we report the final output of the *Parser*, which matches Whitehall and UK Biobank non-imaging variables.

As previously mentioned, both the Parser was a tool specifically developed for this thesis work. The Parser, with its User guide, is already available online to the scientific community on the GitLab platform (https://issues.dpuk.org/eugeneduff/wmh_harmonisation/tree/master/funpack_wmh_bb). Conversely, the extracted harmonised dataset will be available through the Dementias Platform UK. The final dataset, we generated by the Parser and used for the analyses, contains 32 non-imaging variables (see table 4.2) plus 3 imaging variables: total brain volume, total volume of WMH and the percentage volume of lesions with respect to brain volume (WMH%, the independent variable used in our model).

For our extraction, it was necessary to first screen for "visit 2.0" corresponding to the images acquisition phase. In fact, unlike Whitehall, the UK Biobank data are acquired at different time points, depending on the data collection phase. The Parser can be used for a range of variables and converting functions, specified according to the user's needs (as we did in "conversion.py"). Tab. 4.2 summarises the conversions implemented on the 32 variables of interest for the present study.

Whitehall	*Type*	*UK Biobank*	*Type*	*Conversion*
OX.AGE	*Continous, year*	*21003*	*Integer, years*	*Round()*
GenderM1F0	*Categorical*	*31*	*Categorical*	*N/A*
OX. WEIGTH	*Continuous, kg*	*21002*	*Continuous, kg*	*N/A*
OX. HEIGTH	*Continuous, m*	*12144*	*Integer, cm*	*cm → m / 100*
OX.BMI	*Continuous, kg/m^2*	*21001*	*Continuous, kg/m^2*	*N/A*
OX.BP_SYS	*Integer, mmHg*	*4080*	*Integer, mmHg*	*N/A*
OX.BP_DIA	*Integer, mmHg*	*4079*	*Integer, mmHg*	*N/A*
OX.PULSE	*Integer, bpm*	*102*	*Integer, bpm*	*N/A*
ModeratePA	*Continuous, h/week*	*884* *894*	*Integer, day/week* *Integer, min/day*	*(884*894) / 60 = h/week*
VigorousPA:	*Continuous, h/week*	*914*	*Integer, min/day*	*(914*904) / 60 = h/week*

CHAMwalk:	*Continuous, h/week*	*864* *894*	*Integer, day/week* *Integer, min/day*	*(864*874) / 60 = h/week*
TotWalk	*Continuous, h/week*	*894*	*Integer, min/day*	*894/60(min/h) * 7 = = h/week*
SleepDuration	*Continuous, h/day*	*1160*	*Integer, h/day*	*Round()*
HealthClasses	*Categorical 4 classes*	*2178*	*Categorical, 9 classes*	*Clean -1/ -3.* *Convert:* *1 in 4 , 2 in 3* *3 in 2 , 4 in 1*
SmokerStatus_C-NC	*Categorical 0/1*	*20116*	*Categorical* *4 classes*	*Convert:* *0 in 0* *1,2 in 1* *Clean -3*
Cig/day	*Integer*	*3456*	*Integer*	*N/A*
AlcoholStatus_C-NC	*Categorical 0/1*	*20117*	*Categorical 4 classes*	*Convert:* *0 in 0* *1,2 in 1* *Clean -3*
AlcoholU/w	*Continuous, Units/month*	*20403* *20414*	*Categorical 5 classes, U/day* *Categorical 5 classes, day/week*	*- Making them continuous on the mean value* *- 20403*20414 = units/month*
OX.MEDS_TS_ALL	*Integer*	*137*	*Integer*	*N/A*

BNF_CVmedY1N0	*Categorical 0/1*	*6177*	*Categorical 6 classes*	*Binarise Categorical and extract V1*
BPMedRawY1N0	*Categorical 0/1*	*6177*	*Categorical 6 classes*	*Binarise Categorical and extract V2*
BNF_AntideprY1N0	*Categorical 0/1*	*20546*	*Categorical 6 classes*	*Convert: 3 in 1 1,4 in 0 Clean -818*
DiabRawY1N0	*Categorical 0/1*	*2443*	*Categorical 4 classes*	*Clean -1 -3*
CVDRawY1N0	*Categorical 0/1*	*6150*	*Categorical 6 classes*	*Convert: 1,2,4 in 1 -7 in 0 Clean -3*
CESD_depressed	*Categorical 4 classes*	*20510*	*Categorical 5 classes*	*Clean for -818*
OX.EDUC_FT_END	*Integer, years*	*845*	*Integer, years*	*N/A*
HandClasses1R2L3A	*Categorical 3 classes*	*1707*	*Categorical 4 classes*	*Clean for -3*
TMT_A_s	*Integer, s*	*20156*	*Continuous, s*	*Round( )*
TMT_B_s	*Integer, s*	*20157*	*Continuous, s*	*Round( )*
OX.DCOD	*Integer, correct answers*	*20159*	*Integer, correct answers*	*N/A*
OX.DSB	*Integer*	*20240*	*Integer*	*N/A*

Reaction_Time	*Continuous, ms*	*20023*	*Integer, ms*	*Round( )*
WholeBrain	*Integer, mm3*	*25010 (W+G)*	*Integer, mm3*	*(WM+ GM)*
WMH_tot	*Continuous, mm3*	*25781*	*Integer, mm3*	*Round()*
WMH_percent	*Continuous, mm3*	*N/A*	*Continuous, mm3*	*25781 / 25010 *100*

Table 4.2. Parser summary table with conversions rules.

For more details and notes about every step, consult the file "Parser.xlsx" on GitLab.

4.3 Predictive model construction for multi-centre study with prospective harmonisation – Whitehall

The search for the best performing model, able to predict the percentage lesional load WMH% from the non-imaging variables is also a core element in this work, which will be used in Par. 4.4 as further validation of harmonisation addressing the estimate of WMH% from images. In fact, classification statistics (DI, Cluster-level FPR, etc.) could be applied only to the limited manually segmented sets. Conversely, for WMH% validation over the whole populations, our working hypothesis was that a valid image harmonisation should permit well matched prediction models, in the different addressed studies. In other words, we assumed a common epidemiological statistic through the different populations leading to similar relationships between WMH% and its non-imaging correlates. Conversely, different prediction laws would be obtained in the case of biases in WMH% measures, due to poor imaging harmonisation.

As reported in the previous Chapter, all the implemented predictive models (General Linear Model, Random Forest, Ridge Regression and Elastic Net) were built using data obtained in this way: training phase performed according to the following processing options "Single training T1, Rater 2, no biasfield correction, FA still present (*TR1 R2 BF FA)*" and testing phase conducted on the Whitehall dataset. Especially, the resulting optimal model allows us to account for the importance attributed to the variable "Scanner" in predicting WMH%.

This is a key point in the evaluation of the harmonisation results, obtained through the different analysis options introduced (see next section).

Thus, Table 4.3 shows and compares the main indicators of goodness and accuracy of the regressive models implemented.

The considered statistics in Tab. 4.3 are:

- proportion of variance explained by the predictors (R^2);
- root mean square error (RMSE), i.e. quadratic mean of the model residuals;
- Spearman's correlation (r_s) between the actual and predicted value of WMH%.

Model	R^2	*RMSE*	r_s
General Linear (GLM)	*0.14*	*0.24*	*0.46*
Random Forest	*-0.05*	*0.26*	*0.39*
Ridge Regression	*0.16*	*0.24*	*0.47*
Elastic Net	*0.17*	*0.24*	*0.49*

Table 4.3. Performance metrics for the four models implemented with the best resulting one highlighted in blue.

Since our main goal is to maximise the Spearman correlation coefficient, Elastic Net was chosen as the optimal predictive model for the Whitehall study harmonisation (Whll SC1-Whll SC2) and it will be the starting point for the second one extended to BB (Whll-BB, relative results described in section 4.5).

Moreover, Elastic Net was the only one really able to perform an accurate selection of variables among the proposed predictors and it is the best compromise in terms of all the three statistics in Tab. 4.3.

After we obtained the results of the different analysis options on WMH harmonisation in Whitehall, we were able to test the various models also on: (i) Single Training (TR1), Rater 2, FA still present, biasfield correction, Global Threshold (0.9) and (ii) Mixed Training (24+24), FA still present, biasfield correction, Global Threshold (0.9) (see Chapter 3, section 3.8.1.3). Elastic Net reported the best results both before and after harmonisation

obtained through the biasfield correction (i) and the use of a mixed training (ii). In Table 4.4, we can see a review of results in terms of *Spearman Correlation.*

	GLM	*Random Forest*	*Ridge Regression*	*Elastic Net*
T1 R2 BC FA 0.9	*0.36*	*0.31*	*0.37*	*0.40*
24+24 R2 BC FA 0.9	*0.35*	*0.31*	*0.38*	*0.39*

Table 4.4. Spearman Correlation coefficients between actual and predicted WMH% for the two analysis options implemented. Blue boxes indicate the best performing method for each option.

4.4 Evaluation of the influence of different analysis options on WMH harmonisation

In this section, the influence of different analysis options on WMH harmonisation is analysed, when a single option/variable is changed from the optimal setting fixed in through the studies of the previous sections. Namely: (i) rater who performed the manual labelling step, (ii) introduction of the biasfield correction process on images, (iii) composition of the training set used for BIANCA, (iv) presence of the Fractional Anisotropy among the exploited image modalities and (v) choice of the thresholding method. As discussed previously, the goal is to find the best combination able to reduce differences in the WMH measures extracted from the different datasets involved in our study, therefore providing an effect of harmonisation.

Firstly, in section 4.4.1 and 4.4.2. we present BIANCA's performance using all the indicators introduced in the previous chapter and for every tested combination of parameters (for details refer to Table 3.5 and 3.6 respectively). However, we limit to the indicators with highest importance on the final decision: Dice Similarity Index (*DI*) (as a summary measure of overlap) and Cluster-level *FPR* (i.e. we were more interested in achieving high sensitivity to lesion detection). The obtained results, relevant to the different subsets of testing subjects used (12 for *SC1* and 12 for *SC2* in Par. 4.4.1; 12 for *SC1*, 12 for *SC2* and 12 for *BB* in Par. 4.4.2), are graphically shown through *box plots*. For each of the tested combinations of

parameters, we therefore created a plot reporting, on the horizontal axis, the different scanners used and, on the vertical one, the values relative to a specific performance indicator. As an additional way of evaluating the results of the different options, in the second part of this section (4.4.3), the main options are studied vs. age as represented by *scatter plots*. In this context, data are differentiated according to the *Scanner of test*. On the horizontal axis is reported the participants age, while on the vertical one the percentage lesional load (*WMH%*) obtained through the application of BIANCA on the entire datasets. A linear regression line is then fitted to the data to show the correlation between the two variables.

For the same options, we also present the outcome of Elastic Net. As introduced in the previous section, this regression method is able to identify the most predictive variables among the 32 ones included in the original training set. For the three situations of interest we therefore report the most significant non-imaging features, ordered by ascending importance. A reduced importance of the variable "Scanner" was considered as a marker of good harmonisation by the addressed parameter combination.

As before, we split the results in two parts. The first is entirely focused on a between-scanners comparison in the Whitehall dataset, while the second one sees the introduction of UK Biobank in the analysis, which represents case of retrospective harmonisation merging two different datasets.

4.4.1 Multi-centre study with prospective harmonisation - Whitehall

This first section describes the comparison of the obtained BIANCA performance (in terms of both DI and Cluster-level FPR) between *SC1* and *SC2*.

4.4.1.1 Rater

In Fig. 4.2 and 4.3 we present results relevant to the influence of *Rater*. We trained BIANCA using data from SC1 (we exploited 24 manually segmented subjects in one case and 12 in the other) referring to the manual annotation by rater 1, only. FA was included; global thresholding was applied; and biasfield was corrected. In one case (blue) leave-one-out validation was performed on rater 1 annotation (coincident training and validation), while in the latter (orange) validation was performed against rater 2 (changed rater).

The second pair of compared results (orange box in Figure 4.2 and 4.3), were obtained in an analogue way with respect to the previous case, but the manual labelling phase was performed by a different operator, referred to as Rater 2.

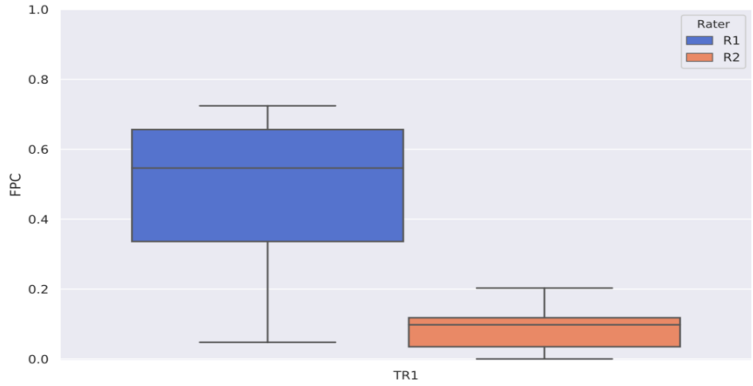


Figure 4.2. Step 1 – Effect of Rater. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask (with leave-one-out) when using masks segmented by rater 1 (R1, blue box) or rater 2 (R2, orange box). Results are relative to SC1, since ratings from two raters was available only for this scanner.

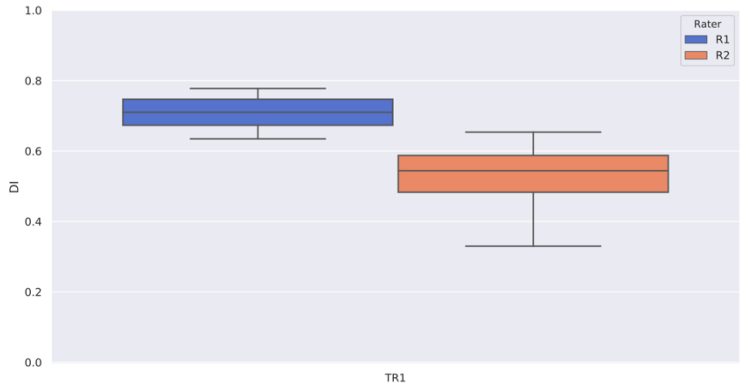


Figure 4.3. Step 1 – Effect of Rater. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask (with leave-one-out) when using masks segmented by rater 1 (R1, blue box) or rater 2 (R2, orange box). Results are relative to SC1, since ratings from two raters was available only for this scanner.

4.4.1.2 Biasfield

In this section we compare a pair of settings for each scanner, in order to assess the influence of *Biasfield* on the resulting BIANCA performance. For the first (left in Figure 4.4 and 4.5), we trained and tested BIANCA using data from SC1 (24 and 12 subjects respectively) and manual masks labelled by Rater 2. The FA scan was included as intensity feature and a global thresholding method was applied. We corrected for the biasfield inhomogeneities in the FLAIR images in one case (green box) and we didn't in the other (blue box). For the second pair of compared settings (right in Figure 4.4 and 4.5), we obtained data in an

analogue way, with the only difference that training and testing phases were performed on subjects imaged by SC2.

The difference in DI between the BF and BC case was statistically significant for SC1 (p-value < 0.001) while not significant for SC2 (p = 0.097).

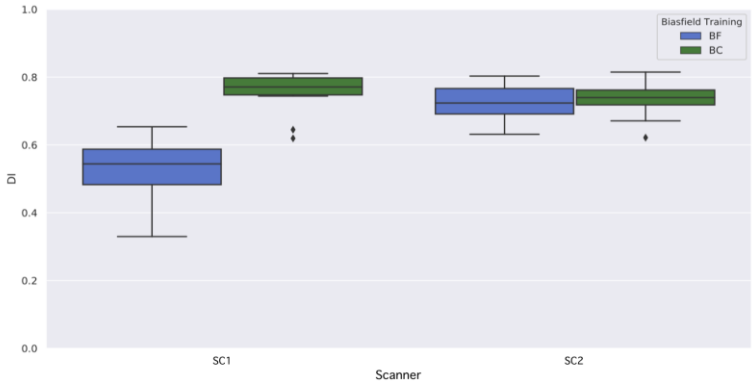


Figure 4.4. Step 2 – Effect of Biasfield correction. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask (with leave-one-out) when correcting for the biasfield inhomogeneities (BC, green box) and when they are still present (BF, blue box). Results are relative to SC1 (left pair of plots) and to SC2 (right pair of plots).

Similarly, Cluster-level FPR between the BF and BC case are significantly different for SC1 (p-value < 0.001) and not significant for SC2 (p-value = 0.259).

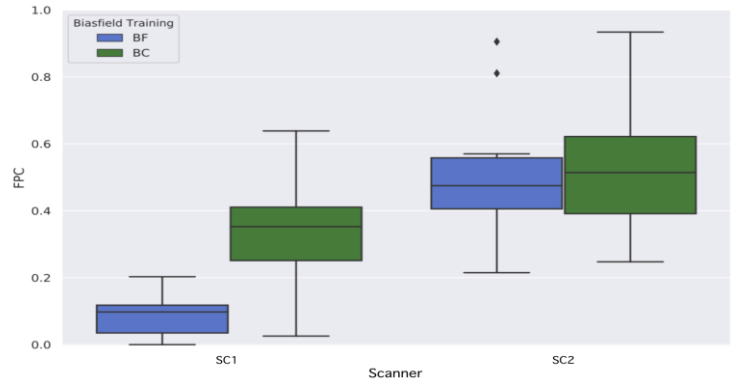


Figure 4.5. Step 2 – Effect of Biasfield correction. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask (with leave-one-out) when correcting for the biasfield inhomogeneities (BC, green box) and when they are still present (BF, blue box). Results are relative to SC1 (left pair of plots) and to SC2 (right pair of plots).

4.4.1.3 Training set

At this point we present results relevant to the effect of different *Training sets*, comparing five settings. We were interested in identifying whether study and scanner specific training sets improved the performance of BIANCA with respect to expert identification of lesions. We trained BIANCA using the subgroups of subjects introduced in the previous chapter: 24 belonging to SC1 (TR1), 24 relative to SC2 (TR2), two training sets obtained as a mix of both scanners (24+24 and 12+12) and an additional reference case in which every scanner was both trained and tested using its own manually segmented masks (*Study specific*). We kept the other processing options fixed: FA included, biasfield corrected and Global thresholding method applied.

Results of DI demonstrate that mixed training sets lead to a more comparable performance across scanners compared to the other training sets used.

Remarkably, they are very close to the case used as reference, represented by the Study specific training set. Obviously, this is the best scoring one, since the training and the testing sets are the same. In particular 24+24 was not significantly different form reference in terms of DI (p-value = 0.392 for SC1), despite a significant difference for SC2 (p-value = 0.014). A similar behaviour characterises Cluster-level FPR, which appears to be lower and more comparable for the 24+24 case, with respect to the others. In particular results relevant to SC2 are significantly close to the study specific ones (p-value = 0.363) thereby confirming the concepts introduced previously. To conclude, results of the 12+12 case also showed a graphical similarity with respect to the reference case and a significant contrast was approached only for the DI case of SC1 (p-value = 0.548).

Since the aim is to harmonise WMH measures across scanners, we performed t-tests comparing the resulting data of SC1 with the ones of SC2, for the most relevant cases involved in our analysis: Study specific (i.e., within both SC1 and SC2, but taken separately) and 24+24.

The difference, for the Study specific case, were significant for the Cluster-level FPR (p-value = 0.013), while not significant for the DI (p-value < 0.001).

On the other hand, for the 24+24 case, even if the performance across scanners was proved to be different for the DI case (p-value = 0.046) the heterogeneity was very low (the obtained p-value is indeed close to the threshold value of 0.05). The performance in terms of Cluster-level FPR, on the other hand, was not significantly different between SC1 and SC2 results.

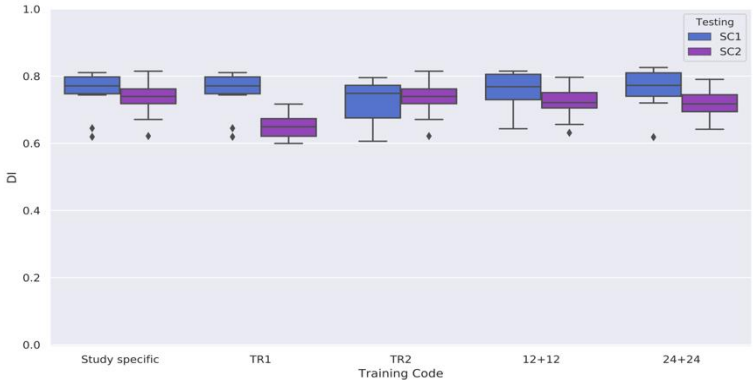


Figure 4.6. Step 3 – Effect of different Training sets. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

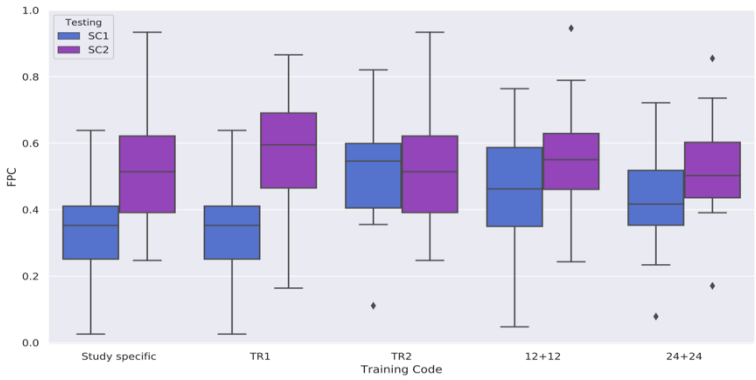


Figure 4.7. Step 3 – Effect of different Training sets. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

4.4.1.4 FA availability

In this section, all previous results are revisited by excluding *FA* as a feature. Outcomes show a significant decrease of performance in terms of DI, with a loss of almost 0.2 (see Figure 4.8 for reference). Conversely, a decrease in Cluster-level FPR was observed in almost all cases, passing from values equal to 0.4 – 0.5 to 0.1 – 0.2 (see Figure 4.9 for reference).

Overall, results confirm that mixed trainings provide the most similar results to the study specific case, particularly for the 24+24 case. The DI p-value was indeed equal to 0.637 for

SC1, in spite of a difference characterising SC2 (p-value < 0.001). An analogue behaviour can be observed for the Cluster-level FPR (SC2 not statistically significant, p-value = 0.199, SC1 statistically different, p-value = 0.006).

More importantly, the mixed 24+24 set was even more comparable across scanners than it was in the previous step, therefore matching the objectives of our work (p-value = 0.462 and p-value = 0.565 for the DI and Cluster-level FPR respectively).

At the light of this result, we concluded that FA removal is likely to decrease the amount of false positive clusters, therefore favouring a higher specificity of the segmentation performance. Moreover, being an MRI modality not very common in clinical contexts, this feature will be excluded from further analysis steps even though it was verified to have a positive effect on the overall segmentation accuracy.

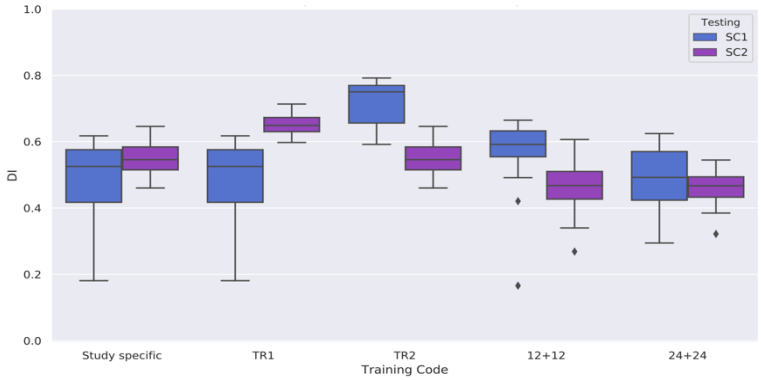


Figure 4.8. Step 4 – Effect of FA exclusion from the training features. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

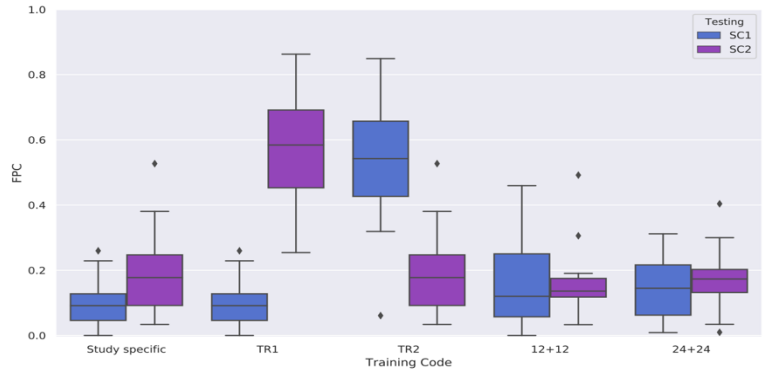


Figure 4.9. Step 4 – Effect of FA exclusion from the training features. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

4.4.1.5 Thresholding

We finally reanalyse all previous results, FA excluded, comparing the global thresholding (applied, so far) with the local thresholding provided by LOCATE. The data shown in Figure 4.9 highlight a significant increase of performance in terms of DI, with an improvement of almost 0.2 when compared to the previous setting. However, the optimised segmentation performance is accompanied by a simultaneous increase in Cluster-level FPR, that results in an overestimation of the total amount of lesions.

The application of a local thresholding method has therefore a dual effect: it compensates for the FA removal from the feature set, re-optimising performance towards the original DI values, at the cost of an increase in the number of voxels incorrectly classified as WMH.

Like in previous steps, results confirmed that mixed trainings provide the most similar results to the study specific case in terms of DI: in the 12+12 case p-values were indeed equal to 0.629 and 0.087, for SC1 and SC2 respectively; in the 24+24 case, on the other hand, results with respect to the study specific training were not significantly different for SC2 (p-value = 0.656), while for SC1 they were slightly different (p-value = 0.043). The Cluster-level FPR gives analogue performances for both comparisons.

Even in this case, as we aim to harmonise WMH measures across scanners, we performed t-tests comparing the results of SC1 and SC2, both before and after the application of LOCATE. Differences were significant for all the different Training sets used, but for the sake of briefness we report here just the most relevant ones: Study specific, 12+12 and 24+24.

For the Study-specific and 12+12 case we obtained p-values < 0.001 for both the DI and the Cluster-level FPR. In an analogue way, the 24+24 case resulted in significantly different performances both in terms of DI (p-value < 0.001 for both SC1 and SC2) and of Cluster-level FPR (p-value < 0.001 for SC1 and p-value = 0.013 for SC2).

The dual effect provided by LOCATE leads us to consider its use as an open question, that will be addressed through further analysis (see Par. 4.4.3).

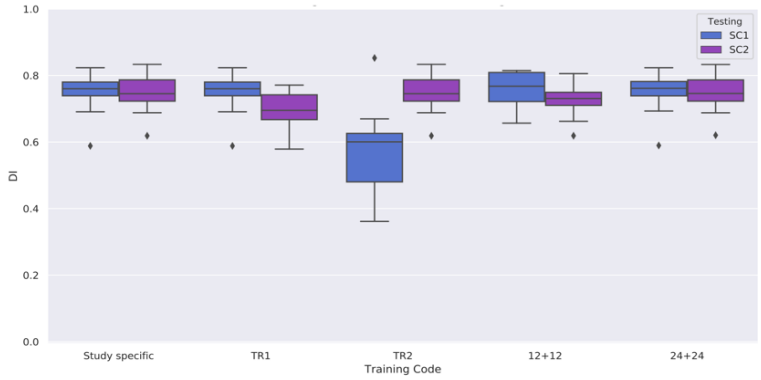


Figure 4.10. Step 5 – Effect of the use of a local thresholding method (LOCATE) on BIANCA output. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

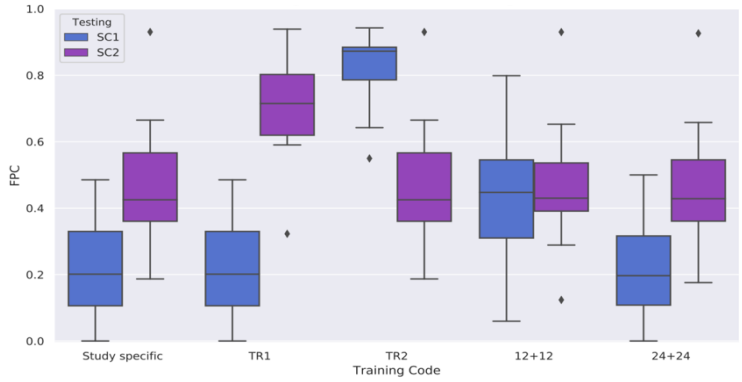


Figure 4.11. Step 5 – Effect of the use of a local thresholding method (LOCATE) on BIANCA output. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask when using different training sets (specified on the x axis): study specific (training with leave-one-out), TR1 (training by 24 subj from SC1), TR2 (training by 24 subj from SC2), 12+12 (training by 12 subj from SC1 and 12 subj from SC2), 24+24 (training by 24 subj from SC1 and 24 subj from SC2). Validation on both SC1 (blue box) and to SC2 (purple box).

4.4.2 Retrospective harmonisation of Whitehall and UK Biobank datasets

As discussed previously, in this second part of the analysis we introduce the UK Biobank dataset. We therefore investigate the influence of different settings of parameters, comparing data relative to three groups of testing subjects: 12 for Scanner 1 (*SC1*), 12 for Scanner 2 (*SC2*) and 12 for the UK Biobank (*BB*). Also in this case, the goal is to find the best combination of parameters able to give comparable performances across datasets, therefore

providing a harmonisation effect. We furthermore try to achieve the best possible performance in terms of segmentation accuracy of the WMH lesions.

4.4.2.1 Training set

We present here the effect of the different *Training sets* tested in the analysis. Since mixed training resulted having the best performance in previous analyses, here we consider only the *12+12* and *24+24* cases, along with the two additional ones, consisting in a combination of the Whitehall and UK Biobank data: *12+12+BB* and *24+24+BB*. All the results presented in Figure 4.12 and 4.13 were obtained using the global thresholding value of 0.9.

From a graphical point of view, it's quite easy to see that *24+24+BB* is the training set able to provide the most comparable performance across datasets. This is noticeable when comparing these results to the ones relevant to the *Study specific* case, as characterised by box plots less similar among datasets.

In order to assess the statistical significance of data comparability for the *24+24+BB* case, we performed a one-way-ANOVA test on the three resulting settings (SC1, SC2 and BB): SC1 and SC2 were not statistically different between each other both in terms of DI and of Cluster-level FPR (p-value = 0.343 and p-value = 0.987 respectively). On the other hand, some differences were observed in the comparisons to the UK Biobank: SC1 vs BB gave a p-value of 0.009 for the DI and of 0.043 for Cluster-level FPR, while SC2 vs BB resulted in a p-value < 0.001 and p-value = 0.030 respectively.

the resulting data where no longer equivalent: SC1 vs BB gave indeed a p-value of 0.009 for the DI and of 0.043 for Cluster-level FPR, while SC2 vs BB resulted in 0.000 and 0.030 respectively.

One-way-ANOVA test relevant to the other investigated Training sets gave similar results with respect to the ones presented above, but the difference between the Whitehall scanners and the UK Biobank are more heterogeneous for both the DI and Cluster-level FPR, being characterised by a difference of almost 0.2 – 0.3 against the 0.1 of the 24+24-BB case. Furthermore, for the 12+12+BB case, the difference between SC1 and SC2 was also significant in terms of DI (p-value = 0.122).

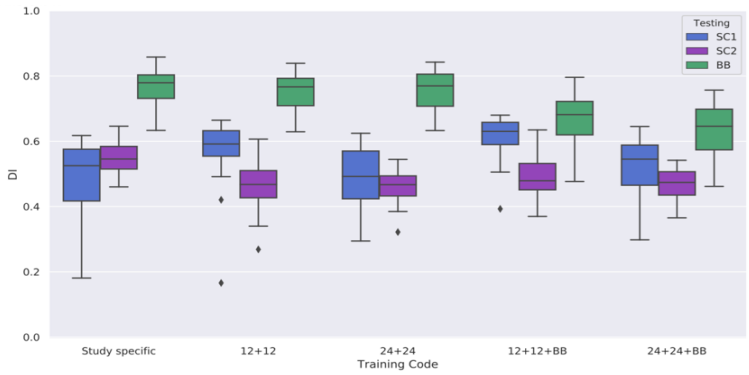


Figure 4.12. Step 6 – Effect of different Training sets. Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

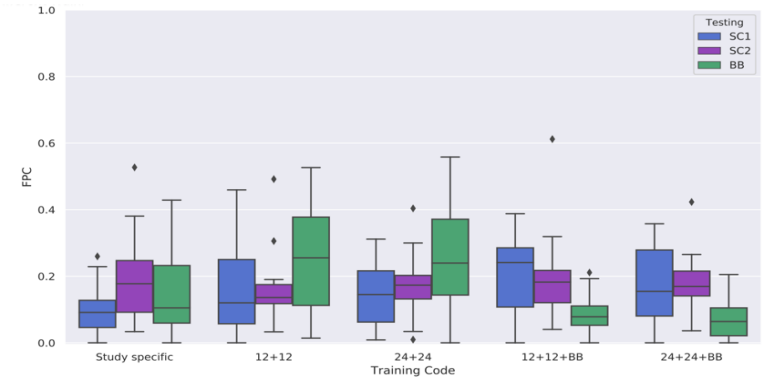


Figure 4.13. Step 6 – Effect of different Training sets. Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

4.4.2.2 Thresholding

In this paragraph we present results obtained after the application of LOCATE to all the options discussed previously. As it happened for the Multi-centre Whitehall study, the automatic tool provided a dual effect: 1) it resulted in a great improvement of the DI performance, with an increase of almost 0.2 – 0.3 for both SC1and SC2 (Figure 4.14). Significance of the difference between the Global and Local thresholding case was assessed

through the use of a t-test, that reported p-values < 0.001 for both the Study specific and 24+24+BB case relative to SC1 and the same ones relative to SC2. The effect was less evident (increase of 0 – 0.1) for the UK Biobank dataset, but still significant (p-value < 0.001 for the Study specific case and p-value = 0.012 for the 24+24+BB one); 2) at the same time, the use of LOCATE resulted in a significant increase in the number of voxels incorrectly classified as WMH. The amount of Cluster-level FPR was indeed much higher than the one obtained with previous settings with an increase of almost 0.2 – 0.4 for all cases (Figure 4.15). Significance of the difference between the Global and Local thresholding was again assessed through the use of a t-test, whose p-values, for the Study specific and 24+24+BB case, were equal to 0.005 and 0.049 for SC1, < 0.001 and equal to 0.002 for SC2 and both < 0.001 for UK Biobank).

Again, the Training set including as many samples as possible from the 3 scanners (24+24+BB) resulted in more comparable performances across the datasets involved in the analysis. Indeed, one-way-ANOVA test showed no significant difference between SC1 and SC2, both in terms of DI and of Cluster-level FPR (p-value = 0.178 and p-value = 0.869 respectively). The same happened for SC2 and UK Biobank with values equal to 0.428 and 0.127 respectively. The only differences were between SC1 and UK Biobank, with p-values equal to 0.011 and 0.044 for DI and Cluster-level FPR.

As the dual effect introduced in the previous section (Par. 4.4.1.5) was confirmed even in this case, LOCATE use remains an open question, requiring further analysis to be addressed (see Par. 4.4.3).

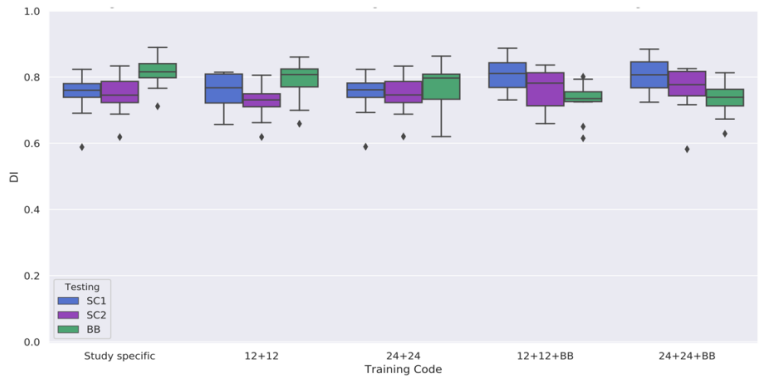


Figure 4.14. Step 7 – Effect of the local thresholding method (LOCATE). Box-plot of the Dice Similarity Index between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj

from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

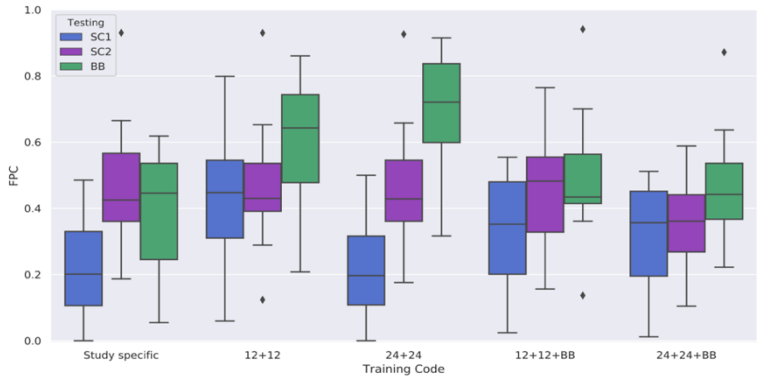


Figure 4.15. Step 7 – Effect of the local thresholding method (LOCATE). Box-plot of the Cluster-level FPR between BIANCA output and the corresponding manual mask for the different training sets used (specified on the x axis): study specific (training with leave-one-out), 12+12 (training by 12 subj from SC1 and 12 from SC2), 24+24 (training by 24 subj from SC1 and 24 from SC2), 12+12+BB (training by 12 subj from SC1, 12 from SC2 and 12 from BB), 24+24+BB (training by 24 subj from SC1, 24 from SC2 and 12 from BB). Validation on SC1 (blue box), SC2 (purple box) and BB (green box).

4.4.3 Evaluation of harmonisation through predictive modelling

Henceforth, we pass to evaluations involving the whole of the considered populations, non-annotated subjects included. Thus, the comparison statistical indexes used so far are no more applicable. To determine whether the above optimisations provide comparable measures of WMH across studies, we conversely utilised predictive modelling from non-imaging variables to determine whether there was a residual scanner/study related bias or a good harmonisation was achieved.

Importantly, this is a core passage of this thesis work aiming at a generalisation of the harmonisation process beyond the limitations of manual annotation due to their high burden and the rater related biases.

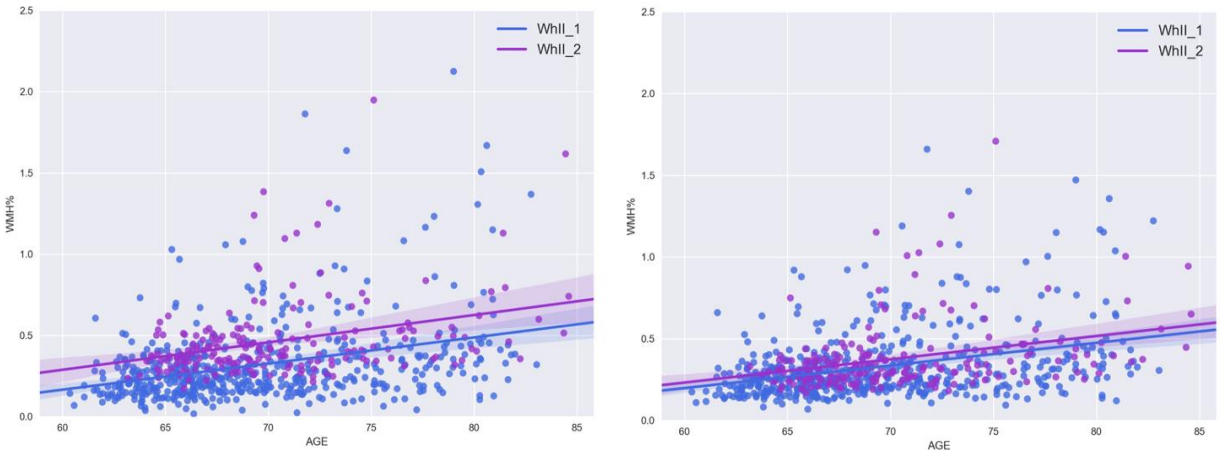
In this section we start by graphical representation of WMH% age dependence for the main comparisons discussed in the previous phase. Remarkably, this time the entire datasets (528 subjects for SC1, 211 for SC2 and 2285 for BB) were addressed. Our aim was to visualise if and how each combination of parameters helped harmonising the WMH volumes (our target measure to harmonise) and if the findings obtained on the subset of subjects with manual masks are generalisable to the entire dataset.

Next, we extended predictive modelling to the entire set of the 32 non-imaging variables by the Elastic Net, in order to assess the relative importance of the different analysis options for predicting WMHs. Specifically, we were interested in how the knowledge of which scanner was used is important to optimise predictions, which is taken as a marker of insufficient harmonisation. For this evaluation we selected only a subset of the tested comparisons, specifically those that showed promising impact on harmonisation: effect of bias field correction, mixed training, and thresholding method.

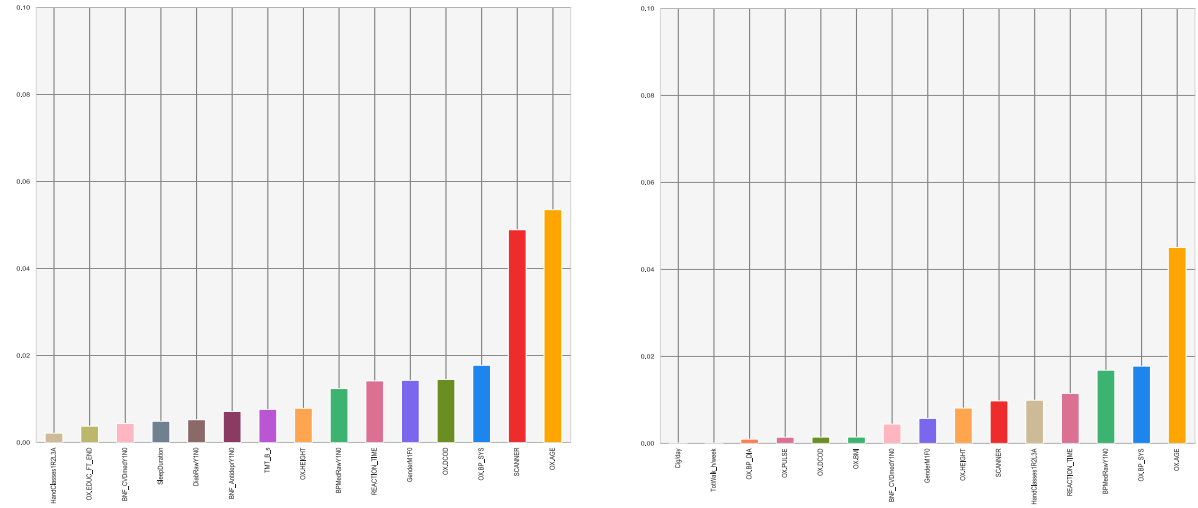
Results are shown relevant to the annotated sub-sets only, in order to limit the density of the scatter-plots and improve their readability. However, we recall that annotation was not used for the predictive modelling, but only for training. Indeed, limiting the modelling assessment to the training cases, permitted us to limit the overall dispersion of results.

The shown settings are the following:

1. Fig. 4.16 – Effect of Bias field correction. *"Single Training (TR1), Rater 2, FA, BF, Global Threshold (0.9)"* versus *"Single Training (TR1), Rater 2, FA, BC, Global Threshold (0.9)"*. Results showed a significant decrease in the volume bias between WMH volumes relative to the different scanners. Moreover, the importance of the variable *Scanner*, assessed through the Elastic Net model, was significantly decreased passing from second to sixth position;



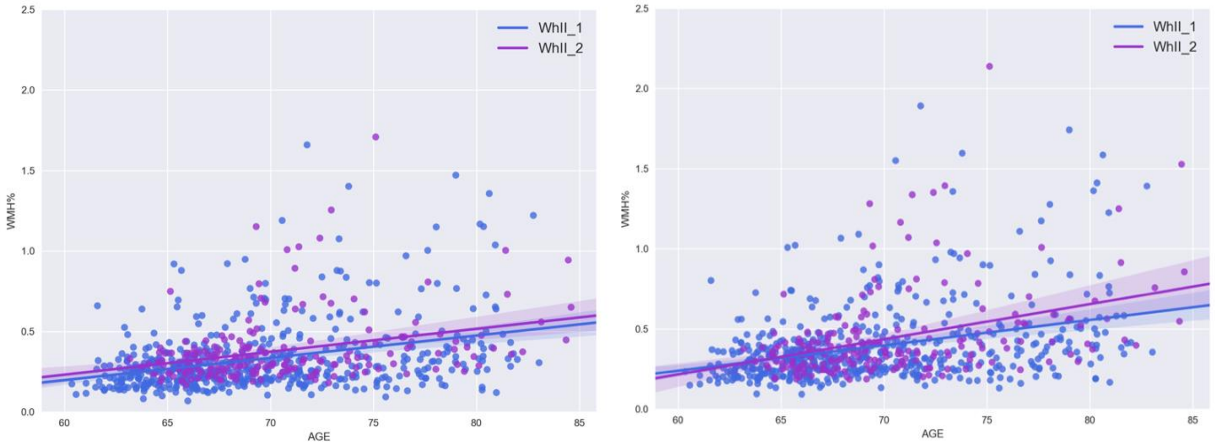
(a) *(b)*



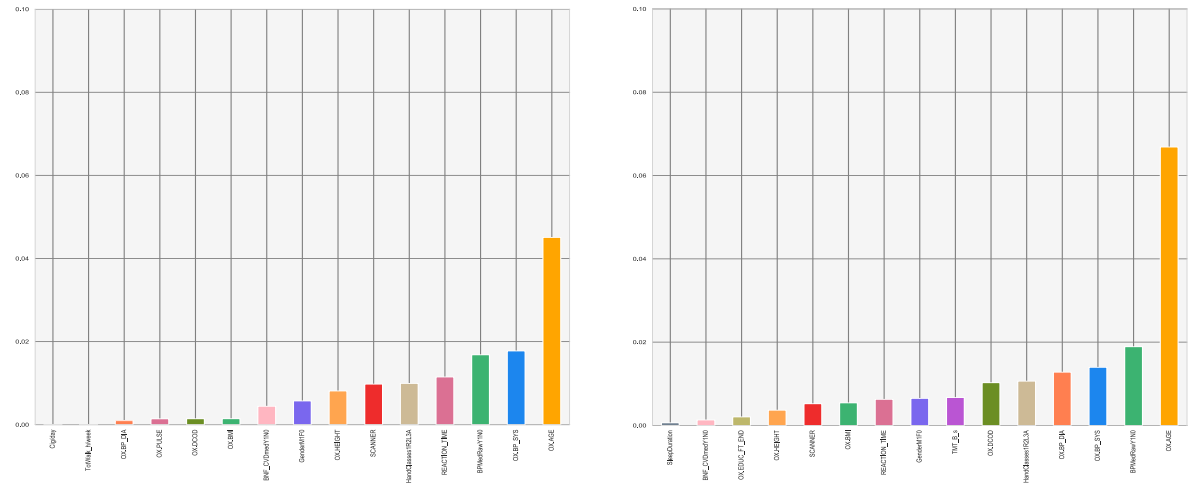
(c) *(d)*

Figure 4.16. Impact of Biasfield correction. Scatter plot relative to the BF (a) and BC case (b), respectively; Importance of the different non-imaging parameters relative to the BF (c) and BC case (d). Bias field correction produces a decrease in importance of the variable "Scanner" (in red) in predicting WMH volume.

2. Fig. 4. 17 – Effect of using mixed vs study-specific training set for BIANCA. *"Single Training (TR1), Rater 2, FA, BC, Global Threshold (0.9)"* versus *"Mixed Training (24+24), Rater 2, FA, BC, Global Threshold (0.9)"*. The volume bias characterising the WMH data was even more decreased compared to the previous setting, despite a slight difference in the slope characterising the regression lines. Furthermore, the importance of the variable *Scanner*, was further decreased, passing from sixth to eleventh position;



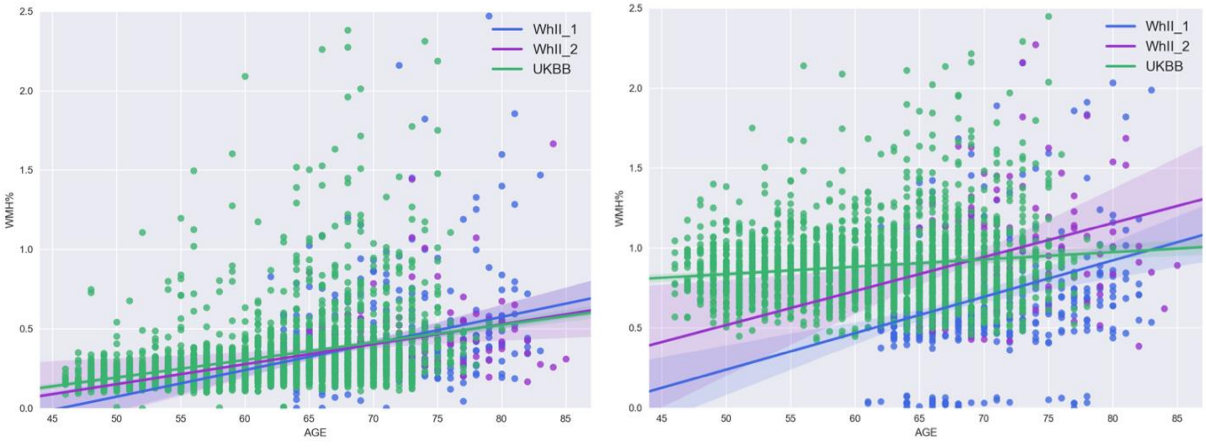
(a) *(b)*



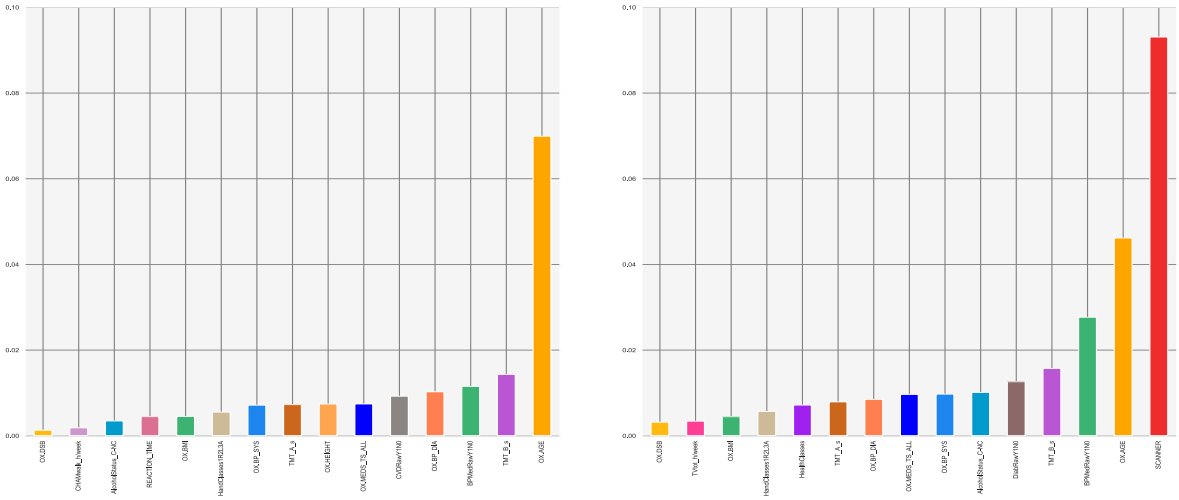
(c) *(d)*

Figure 4.17. Impact of different Training sets. Scatter plot relative to the Single Training (a) and Mixed Training (b), respectively; Importance of the different non-imaging parameters for WMH prediction for the Single Training (c) and Mixed Training case (d). The use of a Mixed Training instead of a single one produces a decrease in importance of the variable "Scanner" (in red) in predicting WMH volume.

3. Fig. 4.18 – Effect of thresholding. *"Mixed Training (24+24+BB), Rater 2, no FA, BC, Global Threshold (0.9)"* versus *"Mixed Training (24+24+BB), Rater 2, no FA, BC, Local Threshold (LOCATE)"*. The volume bias on WMH volumes among datasets was reduced to its minimum using the global thresholding method, while it was still present and even increased when using LOCATE. Furthermore, in the first case the importance of the variable *Scanner*, was not even present in the most predictive features, highlighted by the Elastic Net model. LOCATE, on the other hand, bring its importance to the first position therefore leading us to the decision of avoiding its use in the thresholding procedure.



(a) *(b)*



(c) *(d)*

Figure 4.18. Impact of different Thresholding method. Scatter plot relative to the Global Thresholding (a) and the Local one (LOCATE) (b), respectively; Importance of the different non-imaging parameters for WMH prediction for the Global (c) and Local case (d). The use of a Global Thresholding method produces a decrease in importance of the variable "Scanner" (in red) in predicting WMH volume.

Moreover, since in Whitehall dataset BIANCA's output (WMH%) shows a high correlation with Fazekas visual rating score (r_s = 0.67), BIANCA is considered a good substitute for qualitative evaluation of WMHs, instead of using non-automated visual rating scale which are still frequently used but are time consuming and operator-dependent. (Griffanti et al., 2016)

Thus, to provide whether the standard Fazekas scoring might be more or less affected by a change of scanner we compared the most predictive variables relevant to the model predicting BIANCA's output, with the ones relevant to the model predicting the Fazekas score. We find that the output of the automatic segmentation tool was more independent of the Scanner used (*SC1*, *SC2*) with respect to Fazekas. Indeed, Figure 4.18 highlights the variable *Scanner*, being at second place among the most predictive ones.

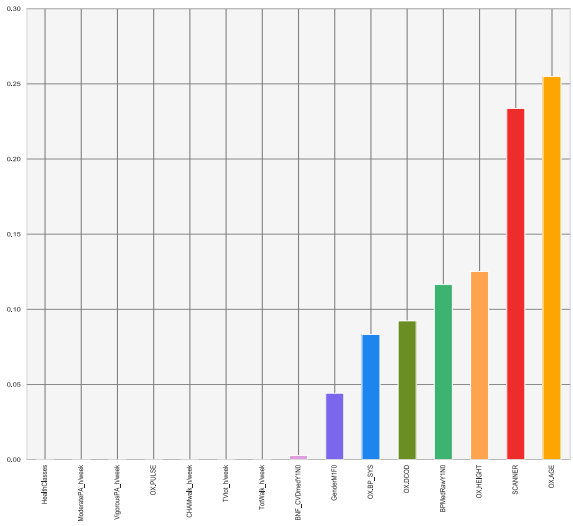


Figure 4.19. Features importance for the prediction of Fazekas score in Whitehall dataset using Elastic Net. This figure has to be compared with Figures 4.16 and 4.17 (c, d).

4.5 Predictive model construction for retrospective harmonisation of Whll and BB

Once we obtained the UK Biobank non-imaging data and turned them into Whitehall format, we were able to apply the various models and validate the consistency of Elastic Net on it testing on the whole datasets, whether annotate or not. The training set was integrated by 12 manually labelled subjects belonging to BB, thus aiming at harmonisation improvement, across datasets.

Accordingly, the final steps of retrospective harmonisation (Training: 24 + 24 + BB, Rater 2, no FA and tested on both Whll and BB) were assessed both in the case of local (LOCATE) and in the global Threshold. All the results in terms of Spearman Correlation r_s are presented in Table 4.5.

	GLM	*Random Forest*	*Ridge Regression*	*Elastic Net*
BB (before harmonisation)	*0.50*	*0.39*	*0.35*	*0.37*
24+24+BB, R2, no FA, Local	*0.28*	*0.25*	*0.28*	*0.29*

24+24+BB, R2, no FA, Global	*0.37*	*0.27*	*0.46*	*0.48*

Table 4.5. Spearman Correlation coefficients between actual and predicted WMH%. (i) model trained on the 12 manually labelled subjects belonging to BB and tested on the whole BB dataset; (ii) model trained on Mixed training 24+24+BB, Rater 2, no FA, local thresholding and (iii) model trained on Mixed training 24+24+BB, Rater 2, no FA, global thresholding, tested on both Whitehall (Whll SC1 and SC2) and UK Biobank (BB). Blue boxes indicate the best performing method for each option.

Using UK Biobank data only, the best result is given by the General linear model, highlighting a more linear relationship between the predictive variables and the WMH%, with respect to the Whitehall one. Otherwise, in both cases characterised by Mixed training it was always Elastic Net that gave the best performances. Then, obtaining the results of the Elastic Net features reduction for both the two datasets, we were finally able to compare them in terms of predictive features and to identify the common variables. Considering the most predictive variables for the best setting for Whitehall data only (training on *Mixed 24+24, Rater 2, FA, Biasfield correction, Global Threshold;* see Figure 4.17 (d)), and those related to UK Biobank (Figure 4.20), it appears that the most predictive variables for WMH% are:

1. Age;
2. Gender;
3. Height;
4. Systolic Blood Pressure;
5. Diastolic Blood Pressure;
6. Body Mass Index (BMI);
7. Blood Pressure Medications;
8. Trail Making TEST_B (TMT-B);
9. Reaction Time;

These nine variables can be divided into demographic (age, gender, height), clinical (Sys/Dia BP and BP meds) and cognitive (TMT-B, Reaction Time) ones.

In particular the TMT-B is part of a common two-part neuropsychological test, in which visuospatial ability (TMT-A) and executive function (TMT-B) are evaluated. Meanwhile Reaction time assesses a person's quickness to react to a stimulus.

The fact that age represents the most significant predictor in not surprising as the significance of age and WMH correlation has been extensively proved by several research studies (e.g. Griffanti et al., 2016; Lee and Preacher, 2013)

Moreover, age shows a significant correlation with WMH% in both the two populations under study (Whll: r_s = 0.35; BB: r_s = 0.51).

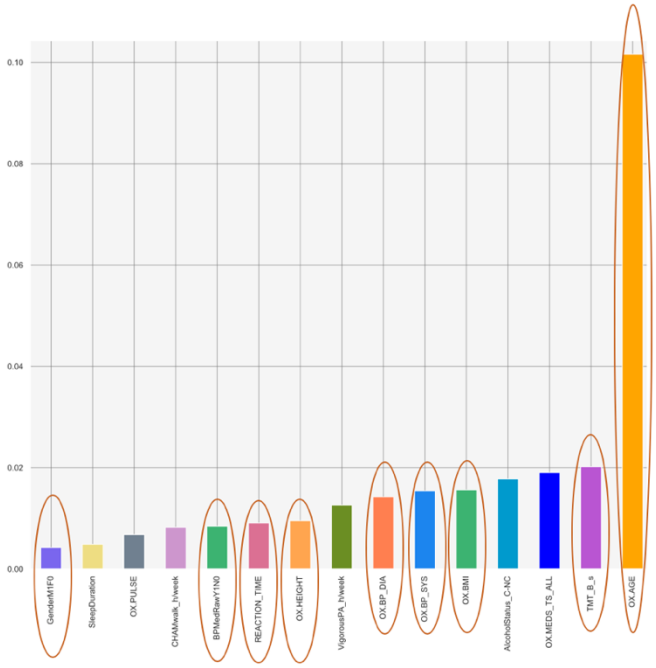


Figure 4.20. Features importance for the prediction of WMH% in UK Biobank dataset using Elastic Net. Red circles represent the most predictive variables common to both Whitehall and UK Biobank dataset.

4.5.1 Gaussian Process

Finally, we explored the use of Gaussian Processes modelling to account for nonlinear patterns in the relationship between our non-imaging variables and WMH%. We were interested in whether predictions trained on a different study (e.g. Whll) could perform to the same level as predictions trained on the target dataset (i.e. UK Biobank).

Initially, we built our Gaussian process regressor with just two variables as predictors and a Radial Basis function as kernel. We applied it separately on Whitehall and UK Biobank, both having WMH% data coming from our best mixed training (*24+24+BB, R2, no FA, Global training*). These basic models allow us to compare the different distributions of WMH volumes with respect to the relative predictive variables. We determined the different

trend of WMH% volume in relation to both age and the main factors of prediction in the two populations. We reported in Figure 4.20 the 3D plot of two main trends analysed.

In both Whitehall and UK Biobank, Age and BMI seem to predict high a high amount of WMH lesion, particularly at their highest values. The same linear pattern characterises the Age and TMT_B pair where, however, the age factor seems to be less determinant. Moreover, if in both Whitehall pairs the surfaces follow the same trend, in UK Biobank surfaces tend to become more and more distant as the values on the y-axis increase. This results into an increase in variance for high values of BMI and TMT_B. respectively.

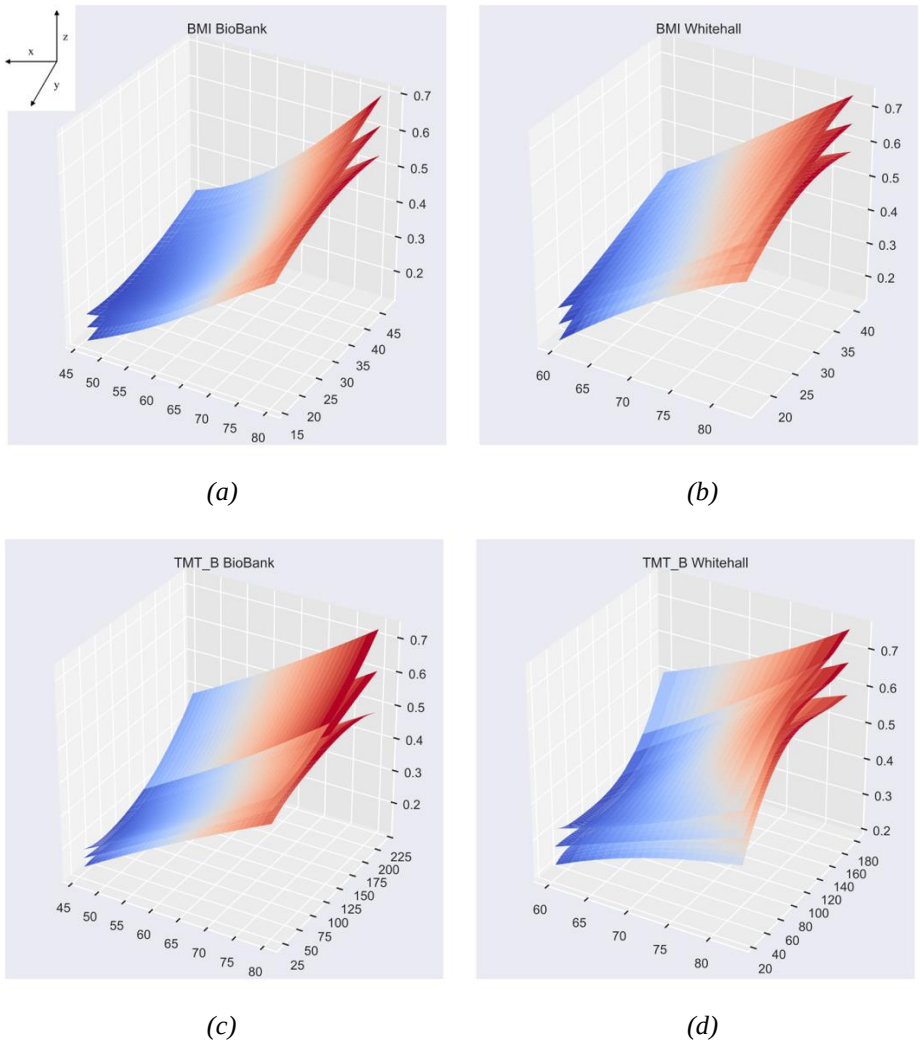


Figure 4.21. 3D plot of WMH% distributions with respect to: Age and BMI (a, b), Age and TMT_B (c. d) in Whitehall (b, d) and UK Biobank (a, c). The three surfaces represent respectively: the mean of predictive distribution for each query points (in the middle) and the mean itself +/- standard deviation (Confidence Interval) above and below it. In all the plots the x axis corresponds to "OX.AGE", the z axis corresponds to "WMH%", while the y axis changes every time depending on the variable analysed.

When we included the nine input variables into the GP's model the model complexity increased. Therefore, we implemented the GP with four different kernel combinations. Those kernels are composed by Radial Basis Function, Matern, Exponential and Rational Quadratic kernels combining with a White noise. Firstly, we fitted the model with UK Biobank data only, which has a greater statistical power given its larger size. Results by the four different kernels where equal in terms of *Coefficient of determination* and *Root-mean-squared deviation* (R^2 = 0.10, RMSE = 0.23), except for the Exponential one (R^2 = 0.02, RMSE = 0.24). Hence, we decided to compare them using the Correlation coefficient between the actual and predicted value and we did the same with Whitehall data, comparing the accuracy of the model on the two different datasets (Table 4.6).

We finally assessed the results of GP model, trained on UK Biobank and tested on Whitehall, compared to the ones obtained both training and testing on Whitehall. This was tested because we are interested in understanding the expected range of WMH% for a new patient regardless of the cohort.

	Trained and tested on BB	*Trained and tested on Whll*	*Trained on BB and tested on Whll*
Radial Basis + White Kernel	*0.43*	*0.38*	*0.37*
ExpSine Squared + WhiteKernel	*0.24*	*0.24*	*0.12*
Rational Quadratic + White Kernel	*0.44*	*0.38*	*0.36*
Matern + White Kernel	*0.43*	*0.36*	*0.37*

Table 4.6. Spearman Correlation coefficients between actual and predicted values of WMH%, obtained by implementing GP using four different combinations of kernels.

As we can see in Tab. 4.6, all of them have almost the same performance in terms of *Correlation* coefficient, except for the Exponential Quadratic one, which are slightly

reduced compared to the ones obtained with linear models (GLM and Elastic Net). Thus, among the implemented kernels, we choose the first one, which shows more comparable results among the different cases. Looking at Figure 4.21 we can appreciate in yellow the WMH% values predicted by our GP's final model trained and tested on UK Biobank data compared to actual BB testing values, while in Figure 4.22 we find our GP's final model trained on UK Biobank and tested on Whitehall compared to actual testing values for both datasets.

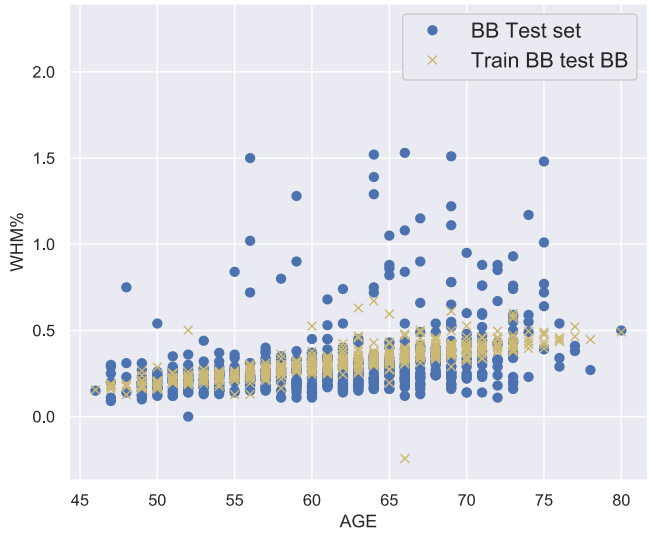


Figure 4.22. Scatterplot of the WMH distribution, obtained by GP trained and tested on UK Biobank, according to age.

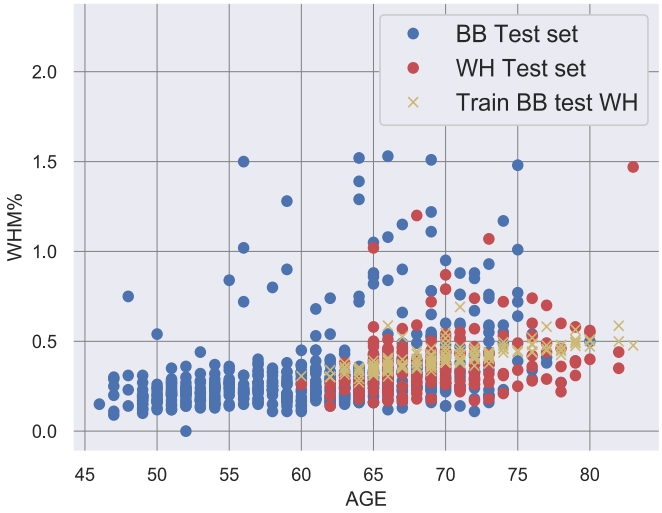


Figure 4.23. Scatterplot of the WMH distribution, obtained by GP trained on UK Biobank and tested on Whitehall, according to age.

More specifically, we investigated the goodness of the previous one with respect to the one trained and tested on Whitehall to verify whether it can be used as a general prediction model to estimate the WMH% for a new patient regardless of the dataset to whom they belong. This will allow us to not re-train the model each time on the specific cohort.

The results show a good overlap of the predicted WMH% volumes with comparable performances (Figure 4.23) and a correlation coefficient of 0.92 between the two.

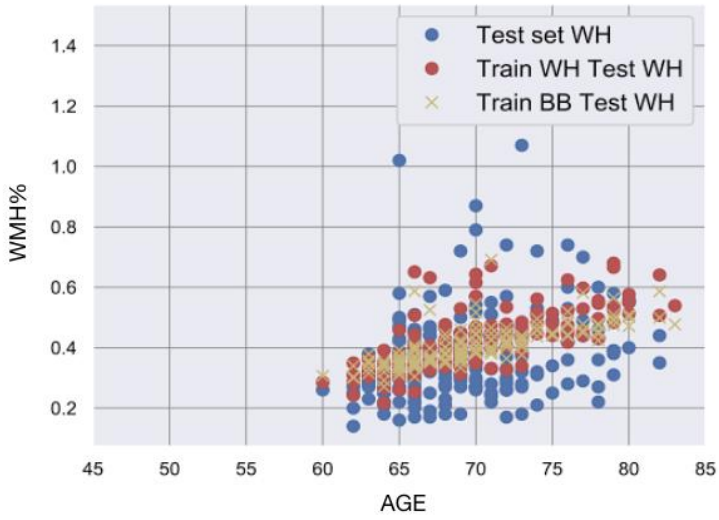
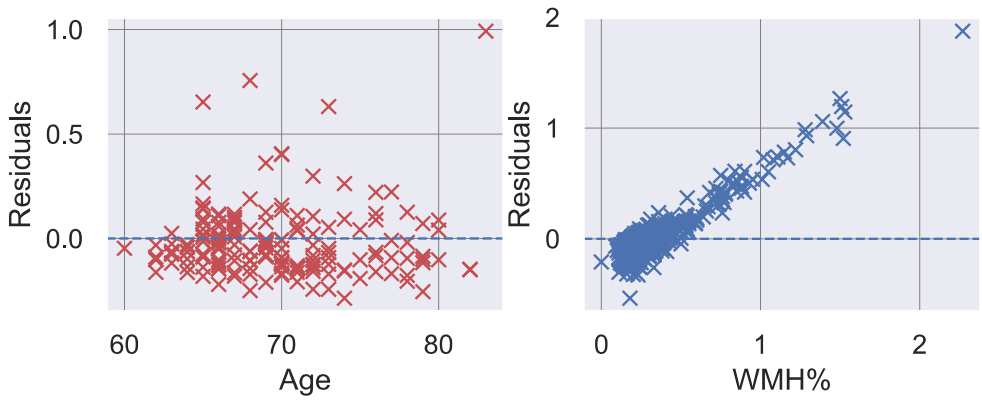


Figure 4.24. Scatterplot of the WMH distribution, obtained by GP trained on UK Biobank and tested on Whitehall (in yellow) vs the one trained and tested on Whitehall (in red).

The residuals analysis also shows similar results. In fact, for both models, the residuals increase considerably as the lesional load increases, following a linear pattern (Figure 4.24 - b, d). While against Age, the residuals appear evenly distributed around 0 with majority of negative values (Figure 4.24 - a, c). In the range between 70 and 80 years we can note a limited number of values due to the lack of data in UK Biobank for that age range.



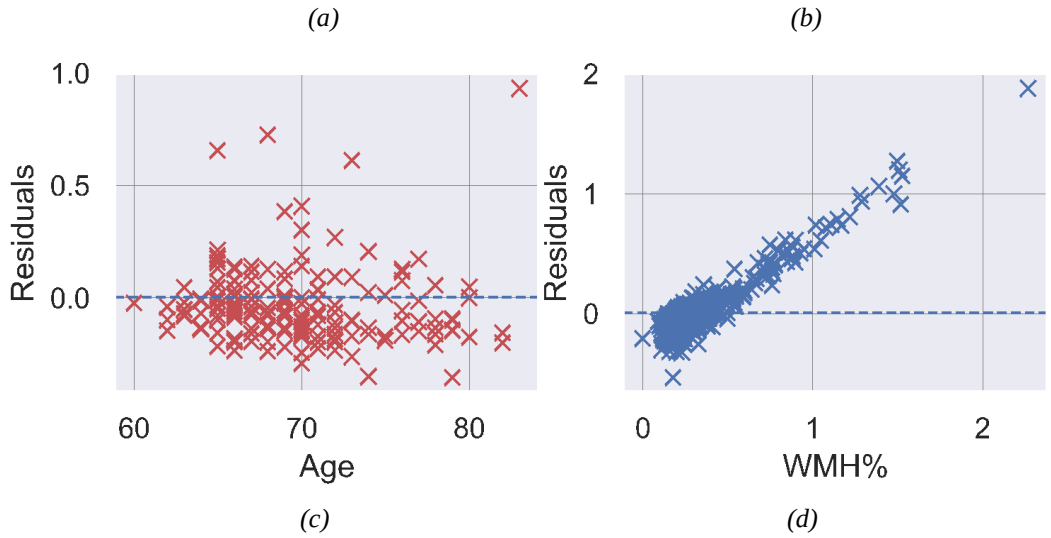


Figure 4.25. Analysis of Model residuals plots: (a) residuals versus age for the model trained and tested on Whitehall, (b) residuals versus WMH% for the model trained and tested on Whitehall, (c) residuals versus age for the model trained on UK Biobank and tested on Whitehall, (d) residuals versus WMH% for the model trained on UK Biobank and tested on Whitehall.

We finally used *The Bland-Altman* (B&A) plot: a graphical method that quantifies the agreement between two methods of clinical measurement plotting their differences vs. their average (Bland and Altman, 1999). In Figure 4.25 we represent the B&A plot of the differences between GP trained and tested on Whll and the one trained on BB and tested on Whll.

Dashed horizontal lines correspond to average differences, and to the limits of agreement (mean ±1.96 std). Ideally the difference between the two methods should be minimal and independent from the quantity measured (WMH%). In our case the difference is acceptable, being within the confidence intervals. However, a linear positive correlation between differences and mean values is also clearly shown, which indicates some residual dependence of errors from values.

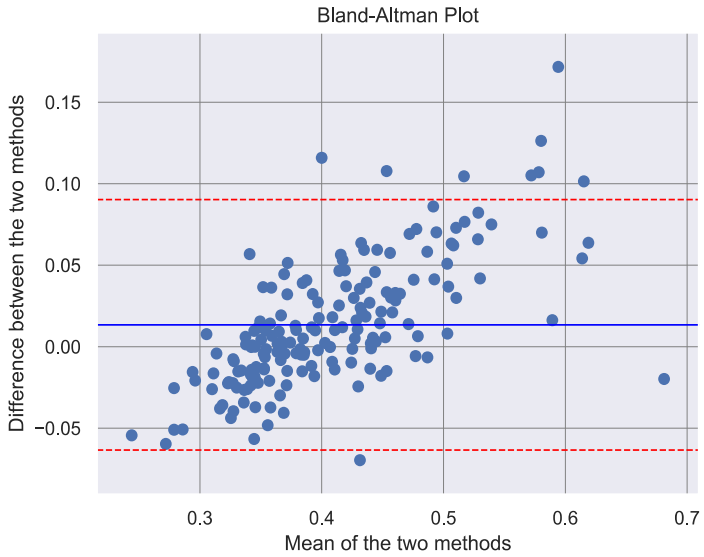


Figure 4.26. Bland Altman Plot of GP's model trained and tested on Whitehall versus the one trained on UK Biobank and tested on Whitehall. The blue horizontal line represents the mean difference between methods, while the red ones are drawn at the limits of agreement, defined as the mean difference ± 1.96 SD of differences.

Chapter 5

“Discussion and Conclusions”

In this Chapter we discuss the results obtained in this work, highlighting the most important aspects that allowed us to achieve our initial goals. We furthermore present the main limitations encountered throughout the development of the analysis, finally outlining briefly some of the possible future developments.

5.1 Discussion

This thesis project focused on the harmonisation of MRI-derived measures of White Matter Hyperintensities (WMHs) across different datasets. Currently, WMHs are intensively studied as an early prognostic sign of a broad band of neurodegenerative pathologies, most of which age-related. This both accounts for the motivation of the present study and also for the challenge of collecting data from diverse imaging datasets. Being able to merge data from different sources has the potential to improve our understanding of the relationship between vascular burden and other clinical and demographic factors with the ultimate aim to inform diagnosis and prognosis of diseases involving vascular lesions (e.g. vascular dementia or stroke). The involved datasets are the UK Biobank (BB) and the Whitehall II imaging study (Whll), two healthy ageing cohorts part of the Dementias Platform UK. However, Whll, represents a multi-centre study gathering data from to a single population, acquired with the same acquisition protocol but exploiting two different MRI Scanners (SC1: 3T Siemens Verio, SC2: 3T Siemens Prisma). On the other hand, the BB includes data from a different population, imaged using a third Scanner (SC3: 3T Siemens Skyra) and a different acquisition protocol. For this reason, our work was divided in two separate parts, relevant to different scenarios: the first is a retrospective harmonisation across scanners (Whll SC1 vs Whll SC2) added to a pre-existing prospective one, given by the Whll study design; the second, on the other hand, is again a retrospective integration process, but it challenges the problem of harmonising two completely independent studies; i.e., Whll and BB.

In order to compensate for the lack of prospective harmonisation between Whll and BB also in the collection of non-imaging variables (e.g., smoking habits, cognitive tests, alcohol consumption, etc), the first part of our work focused on their retrospective harmonisation. Thus, we created a specific “Parser” tool for the conversion of the BB data into the Whll format, to remove differences caused by heterogeneous data collection protocols and different units of measurement. The above-mentioned Parser was successfully implemented and uploaded online on the GitLab platform, as an open source tool available for future analysis using BB or other datasets that would require non-imaging data harmonisation, since it is fully customisable.

Once obtained homogeneous non-imaging data across studies, we focused on using them to predict the volumetric amount of WMH lesions (as percentage of total brain volume, WMH%). The general aim was to find a model able to account for the variability related to

demographic and clinical characteristics of the individuals and, moreover, to evaluate the relationship between WMH% and its major correlates or risk factors. To pursue this goal, we investigated which mathematical model could better fit the non-imaging variables that were available in our datasets and identified from the literature as potentially related to WMH%. Among the four models we tested (GLM, Random forest, Ridge regression and Elastic Net), Elastic Net gave the best result, providing an optimal compromise across the R^2, RMSE and Spearman's correlation values r_S. This model was also the only one able to perform feature selection among the involved non-imaging data, allowing us to evaluate how important the *Scanner* variable was in predicting the volumetric amount of lesions. We therefore used Elastic Net for two aims: firstly, as a metric to judge the success of imaging data harmonisation (i.e., the used scanner should have little or no predictive value if the measures are well harmonised) on the whole datasets, since the modelling approach was applicable also to the non-annotated data. Secondly, to identify the most predictive clinical and demographic variables common between the two datasets, to include in a more sophisticated model for WMH% prediction.

At this point, we aimed at identifying processing choices that produced well-matched measures of WMHs, the imaging variable of interest, across the different data. We did that by testing specific pre-processing techniques (such as biasfield correction) and optimising the robustness of BIANCA, the automatic lesion segmentation tool used in our work, based on trained (alias, supervised) k-NN clustering. The latter goal was pursued by exploring the effect of the rater who generates the training data, finding a general training set able to gain comparable performances across different datasets, studying the effect of excluding FA as intensity feature, and exploring the effect of different options for thresholding the lesion probability map obtained as output from BIANCA.

As BIANCA uses several different options, we conducted a preliminary analysis to assess the best combination of options to obtain comparable outcomes between the two Whll scanners in terms of segmentation accuracy. This ensured that no further bias was being introduced by the parameter choices. The best results were found to be the default parameters, found by Griffanti and colleagues (Griffanti et al., 2016) during their work: number of *Training points* equal to 2˙000, number of *Non-lesion points* equal to 10˙000, a *Patch size* of 3 and a *Spatial Weight* of 2.

While the preliminary part of the analysis was mainly focused on the optimisation of basic BIANCA options, the second part aimed at assessing the influence of specific parameters as to their impact on harmonisation. Results of the step-by-step analysis, described during the previous chapters and implemented in order to assess their impact on the ultimate WMH segmentation performance, are summarised and discussed as follows:

- Regarding the impact of the Rater who generated the WMH masks used to train the supervised algorithm, performance indicators such as DI and Cluster-level FPR, showed that this parameter greatly influenced the segmentation performance. In our study we used masks from the same rater whenever possible (Whll SC1, SC2), but this is not feasible in the large scale (already when adding BB data). This outcome highlights the differences among expert radiologists (alias, raters) in charge of manual labelling and also the need to standardise the definition of WMH, especially if automated supervised tools are planned to be used.
- In terms of performance indicators, biasfield correction proved to greatly improve results by correcting inhomogeneities of the radiofrequency (RF) field. Indeed, the strongest impact on segmentation performance was registered for those images slightly affected by intensity variations across space. Scatter plots comparing the BF (Biasfield not removed) and BC (Biasfield Corrected) case showed a significant decrease in the WMH volume bias between scanners. The accomplishment of a harmonising effect was confirmed by a great decrease in the importance of the *Scanner* variable, assessed through the Elastic Net model. Therefore, in conclusion, it's always recommendable to correct MRIs from biasfield;
- The use of a mixed training set, combining images from all of the datasets involved in our analysis, helped obtaining better and more comparable outcomes in terms of DI and Cluster-level FPR. This was true for both the above-mentioned scenarios (Whll SC1 vs Whll SC2; Whll vs BB) and was moreover proved by a further decrease in the volume bias when comparing outputs of the mixed training case with respect to the single training ones. Furthermore, in both settings (i – *Mixed Training 24+24, Rater 2, FA, BC, Global Threshold); and ii – Mixed Training 24+24+BB, Rater 2, no FA, BC, Global Threshold)* the Elastic Net model highlighted a significant decrease in the importance of the variable *Scanner*, that was no longer among the most predictive ones. These finding suggests that having enough examples of

heterogeneous datasets to use for the training phase, the automatic segmentation tool could generalise well also to unseen data beyond the training set.

- The exclusion of Fractional Anisotropy (FA) from the exploited MRI modalities, resulted in a significant decrease of the DI performance with respect to the previous setting. However, this was accompanied by a corresponding decrease in the amount of Cluster-level FPR, that proved a substantial reduction in the overall level of lesion overestimation. The effect was therefore dual: a significant worsening of the overall accuracy but, at the same time, an increase in the specificity characterising the segmentation performance. The choice relative to the its inclusion/exclusion is therefore strongly dependent on the nature of the problem that needs to be addressed. Since FA is a time-consuming MRI modality, not very common in clinical contexts, and our goal was to obtain a general pipeline, combining information from different cohorts, we decided to remove it from the analysis. This is in line with the BB imaging analysis team, who decided to extract WMHs by T1 and FLAIR only, even if FA was available in most cases.
- Finally, assessing the influence of thresholding technique, we found out that LOCATE (Locally Adaptive Threshold Estimation) actually provided a dual effect: it resulted in a great improvement of the DI performance, but, at the same time, significantly increased the number of voxels incorrectly classified as WMH. This happened for both the Whll SC1 vs Whll SC2 case and the Whll vs BB one. Additional evidence was provided by the scatter plots of WMH% vs. Age, which showed a significant increase in the volume bias mentioned above. Even though this result seems in contradiction with the increase in segmentation accuracy (DI), this ambiguity might be explained through the following hypothesis: LOCATE works well on data characterised by high lesional loads, therefore providing high values of the performance indicator for those subjects belonging to the training set (which were selected among high lesional load participants because the lesions are better defined and easier to be manually segment and also because BIANCA showed to better results when trained on subjects with high lesional loads - Griffanti et al., 2016) . On the other hand, when it is applied to the whole population, characterised by a more variable WMH%, often low, the automatic tool is no longer able to provide a good segmentation performance. We therefore conclude that the choice of the thresholding

method is strongly dependent on the nature of the problem and decided to exclude LOCATE in our harmonisation setting, in the sake of generality.

In summary, among all the settings discussed above, the one offering the best trade-off among performance indicators, reduction in the volume bias, low effect of the used scanner, and generality of application is given by the following combination of parameters: *Mixed Training (i.e., in this context Whll 24+24 with Rater 2, plus BB), no FA, BC, Global Threshold*.

Even though FA and LOCATE resulted being important factors in the achievement of increased BIANCA performances, they were no longer helpful in the context of retrospective harmonisation, when trying to integrate the heterogeneous datasets involved within our study (Whll SC1, Whll SC2, BB). For this reason, we excluded FA from our harmonisation setting and we exploited a global thresholding.

Results showed comparable and accurate outcomes both in terms of DI and Cluster-level FPR. The resulting scatter plot displayed very uniform and integrated data, almost overlapping between each other. Furthermore, the variable *Scanner*, was no longer found among the most predictive variables selected by Elastic Net model. These results, all together, proved the accomplishment of a robust harmonisation effect on the different datasets involved in our study, that were eventually well integrated and compatible.

In addition, the major non-imaging variables, age and cognitive functions, selected by our modelling are in keeping with the main WMH correlates reported in the literature. In fact, age has been highly related to the presence of WMHs in literature (Simoni et al., 2012) and their correlation has been extensively proved by several research studies (e.g., Griffanti et al., 2016; Lee and Preacher, 2013). Other studies (e.g., Kim et al., 2008) have shown that presence and severity of WMH are consistently related to cognitive function and cognitive test scores in the elderly population. Moreover, the prevalence of WMH increases with increasing vascular risk factors, including hypertension (Dufouil et al., 2001; Maillard et al., 2012) and aggressive blood pressure reduction (Sabayan B et al., 2013).

Once obtained the optimal analysis option setting (*Mixed Training, no FA, BC, Global Threshold*) we tried to improve our WMH prediction model, both in terms of input data (i.e. using the harmonised WMH% volumes) and model complexity. After defining Elastic Net as the best linear regression model (compared to GLM, Random forest and Ridge Regression), we trialled Gaussian Process regression (GP), a technique that can describe

non-linear relationships and can provide confidence intervals on WMH% predictions. We fitted the GP model with the non-imaging variables that best predicted WMH% using Elastic Net and that were common to both Whll and BB, and we found the optimal hyper-parameters, tuned specifically on our datasets.

When trained and tested on BB, GP produced satisfying results (Spearman correlation between WMH% actual and predicted value: r_s= 0.43), considering that the involved predictive variables were only nine. However, the non-linear GP predictor did not outperform the linear Elastic Net one. Next, the model was assessed for its ability to predict in new data (Whll), by comparing models: (1) trained on BB and tested on Whll and (2) both trained and tested on Whll. This allows us to verify whether (1) GP can be used as a general WMH% prediction model regardless of the employed scanner and dataset. In this way any deviation from the predicted value (used as normative value) can be clinically interpreted as potentially pathological.

Results gave greatly comparable performances in terms of model-specific Spearman correlations (1. r_s= 0.37, 2. r_s= 0.38), a high correlation value between each other (r_s= 0.92) and very similar residual analysis plots. Moreover, the use of a *Bland-Altman* plot confirmed a good agreement between the two models. However, a linear pattern between relative errors and WMH%, yet within the confidence interval, was observed. It can be explained by the low amount of data within BB of subjects over 70 years, usually associated with greater lesional loads, which is likely to bias the BB trained GP model.

In summary, it was evident that the GP model trained on BB was able to give satisfying outputs either being tested on BB itself or on Whll. For this reason, it can be used as a general prediction model to estimate the WMH% for a new patient regardless of the cohort to whom it belongs. This was an additional proof of the effect of data integration achieved throughout our work.

5.2 Conclusions

With this work we achieved a successful harmonisation of measures of White Matter Hyperintensities (WMHs) of presumed vascular origin across different large datasets. Therefore, we contributed overcoming the barriers that hinder the possibility by the clinical and the scientific communities to integrate observations across studies. Our work led to three main outcomes: firstly, we developed a Parser to harmonise non-imaging variables across

studies, which is specific for the datasets used in our work but can be adapted to other studies. Secondly, we optimised the analysis pipeline for extracting comparable WMH measures across scanners/studies, including increasing the accuracy and robustness of pre-existing software (BIANCA), used to segment white matter lesions. Finally, we proposed a model to predict the amount of WMHs from demographic and clinical data and also to assess harmonisation beyond the availability of annotated sets. As a future clinical application, the predicted value obtained from a normative population can be compared with the measured value to help detecting deviations from the norm. This has the potential to provide physicians with a tool that could help diagnosis and prognosis for diseases involving WMHs, like vascular dementia, cognitive impairment and stroke.

Further future developments may address the automatic WMH segmentation tool BIANCA. Our work tried to optimise both segmentation accuracy and sensitivity. However, the amount of false positive clusters could be reduced introducing a global thresholding value on the minimum cluster size classifiable as lesion, therefore obtaining an overall increase in the precision of segmentation.

Other developments would improve the harmonisation method validation trying to integrate additional cohorts with the ones considered so far, in order to test whether the general harmonising effect, achieved throughout our work, would still be valid on completely new and unseen datasets.

Preliminary results of the analysis carried out during this project have been or will be presented at international conferences by the following accepted works: 1) "Harmonising white matter hyperintensities measures across studies: impact of BIANCA training options", L. Griffanti, **I. Bertani**, **V. Bordin**, I. Mattioli, G. Zamboni, S.Suri, E. Zsoldos, K.P. Ebmeier, M.M. Laganà, G. Baselli, M. Jenkinson, C.E. Mackay, E. Duff. Organisation for Human Brain Mapping Conference, Rome 2019; 2) "Between- and within-rater agreement in white matter hyperintensity segmentation from manual rating and a supervised automated classifier, FSL-BIANCA", L. Griffanti, I. Mattioli, **V. Bordin**, **I. Bertani**, S. Suri, E. Zsoldos, K.P. Ebmeier, C.E. Mackay, G. Zamboni. Accepted as research presentation at the European Congress of Radiology, Vienna 2020.

A journal paper is in preparation.

Bibliography

[1] "What are White Matter Hyperintensities Made of? Relevance to Vascular Cognitive Impairment", Joanna M. Wardlaw, Maria C. Valdes Hernandez et Susana Muñoz-Maniega, 2014.

[2] "BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities", Ludovica Griffanti, Giovanna Zamboni, Aamira Khan, Linxin Li, Guendalina Bonifacio, Vaanathi Sundaresan, Ursula G. Schulz, Wilhelm Kuker, Marco Battaglini, Peter M. Rothwell, Mark Jenkinson, 2016.

[3] "Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis", Lauren Griffith, Edwin van den Heuvel, Isabel Fortier, Scott Hofer, Parminder Raina, Nazmul Sohel, Hélène Payette, Christina Wolfson Sylvie Belleville, 2013.

[4] "Inter-site and inter-scanner diffusion MRI data harmonization", H. Mirzaalian, L. Ning, P. Savadjiev, O. Pasternak, S. Bouix, O. Michailovich, G. Grant, C.E Marx, R.A. Morey, L.A. Flashman, M.S. George, T.W. McAllister, N. Andaluz, L. Shutter, R. Coimbra, R.D. Zafonte, M.J. Coleman, M. Kubicki, C.F. Westin, M.B. Stein, M.E. Shenton, and Y. Rathi, 2016.

[5] "Grading and interpretation of white matter hyperintensities using statistical maps", Wi-Sun Ryu, Sung-Ho Woo, Dawid Schellingerhout, Moo K. Chung, Chi Kyung Kim, Min Uk Jang, Kyoung-Jong Park, Keun-Sik Hong, Sang-Wuk Jeong, Jeong-Yong Na, Ki-Hyun Cho, Joon-Tae Kim, Beom Joon Kim, Moon-Ku Han, Jun Lee, Jae-Kwan Cha, Dae-Hyun Kim, Soo Joo Lee, Youngchai Ko, Yong-Jin Cho, Byung-Chul Lee, Kyung-Ho Yu, Mi-Sun Oh, Jong-Moo Park, Kyusik Kang, Kyung Bok Lee, Tai Hwan Park, Juneyoung Lee, Heung-Kook Choi, Kiwon Lee, Hee-Joon Bae, Dong-Eog Kim, 2014.

[6] “White Matter Disease as a Biomarker for Long-Term Cerebrovascular Disease and Dementia”, Aurauma Chutinet, Natalia S. Rost., 2014.

[7] “Age- and sex-specific rates of leukoaraiosis in TIA and stroke patients”, Michela Simoni, Linxin Li, Nicola L.M. Paul, Basil E. Gruter, Ursula G. Schulz, Wilhelm Küker, Peter M. Rothwell, 2012.

[8] “Handbook of MRI Technique”, Catherine Westbrook, 2008.

[9] “Diffusion Tensor Imaging of the Brain”, Andrew L. Alexander, Jee Eun Lee, Mariana Lazar Aaron S. Field, 2007.

[10] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/.

[11] “Multi-Modality Imaging: Applications and Computational Techniques”, Mauren Abreu de Souza, Humberto Remigio Gamba, Helio Pedrini, 2018.

[12] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT.

[13] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FNIRT.

[14] “Segmentation of Brain MR Images Through a Hidden Markov Random Field Mod el and the Expectation-Maximization Algorithm”, Yongyue Zhang, Michael Brady, and Stephen Smith, 2001.

[15] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FAST.

[16] “Review of brain MRI image segmentation methods”, M. A. Balafar, A. R. Ramli, M. I. Saripan, S. Mashohor, 2010.

[17] “Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review”, Maria Eugenia Caligiuri, Paolo Perrotta, Antonio Augimeri, Federico Rocca, Aldo Quattrone, Andrea Cherubini, 2015.

[18] “Probabilistic segmentation of white matter lesions in MR imaging”, Petronella Anbeek, Koen L. Vincken, Matthias J.P. van Osch, Robertus H.C. Bisschops, and Jeroen van der Grond, 2004.

[19] “Segmentation of age-related white matter changes in a clinical multi-center study”, Tim B. Dyrby, Egill Rostrup, William F.C. Baaré, Elisabeth C.W. van Straaten,

Frederik Barkhof, Hugo Vrenken, Stefan Ropele Reinhold Schmidt, Timo Erkinjuntti, Lars-Olof Wahlund, Leonardo Pantoni, Domenico Inzitari, Olaf B. Paulson, Lars Kai Hansen, Gunhild Waldemar, on behalf of the LADIS study group, 2008.

[20] "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly", F. Admiraal-Behloul, DMJ. van den Heuvel, H. Olofsen, MJP. van Osch, J. van der Grond, MA. van Buchem, JHC. Reiber, 2005.

[21] "White matter lesion extension to automatic brain tissue segmentation on MRI", Renske de Boer, Henri A. Vrooman, Fedde van der Lijn, Meike W. Vernooij, M. Arfan Ikram, Aad van der Lugt, Monique M.B. Breteler, Wiro J. Niessen, 2009.

[22] "The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults", C. DeCarli, D.G.M. Murphy, M. Tranh, C. L. Grady, J. V. Haxby, J. A. Gillette, J. A. Salerno, A. Gonzales-Aviles, B. Honvitz, S. I. Rapoport, M. B. Schapiro,1995.

[23] "White matter hyperintensity burden in elderly cohort studies. The Sunnybrook Dementia Study, Alzheimer Disease Neuroimaging Initiative, and Three-City Study", Joel Ramirez, Alicia A. McNeely, Christopher J. M. Scott, Mario Masellis, Sandra E. Black, 2015.

[24] "Machine Learning", Tom M. Mitchell, 1997.

[25] "Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding", Vaanathi Sundaresan, Giovanna Zamboni, Campbell Le Heron, Peter M. Rothwell, Masud Husain, Marco Battaglini, Nicola De Stefano, Mark Jenkinson, Ludovica Griffanti, 2018.

[26] "Handbook of computational Geometry", J.-R. Sack, J. Urrutia, 2000).

[27] "Random Forests", Leo Breiman, 2001.

[28] "Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: Comparing meta and mega analytical approaches for data pooling", Kochunov P., Jahanshad N., Sprooten E., Nichols T., Mandl RC., Almasy L., Booth

T., Brouwer RM., Curran JE., de Zubicaray GI., Dimitrova R., Duggirala R., Fox PT., Hong LE., Landman BA., Lemaitre H., Lopez LM., Martin NG., McMahon KL., Mitchell BD., Olvera RL., Peterson CP., Starr JM., Sussmann JE., Toga AW., Wardlaw JM., Wright MJ., Wright SN11, Bastin ME., McIntosh AM., Boomsma DI., Kahn RS., den Braber A., de Geus EJ., Deary IJ., Hulshoff Pol HE., Williamson DE., Blangero J., van 't Ent D., Thompson PM., Glahn DC., 2014.

[29] "Harmonizing diffusion mri data across multiple sites and scanners", Mirzaalian, H., Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M.,Westin, C.F., Shenton, M.E., Rathi Y., 2015.

[30] "Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters", Cetin Karayumak S., Bouix S., Ning L., James A., Crow T., Shenton M., Kubicki M., Rathi Y., 2018.

[31] "Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome Longitudinal Study", Forsyth JK., McEwen SC., Gee DG., Bearden CE., Addington J., Goodyear B., Cadenhead KS., Mirzakhanian H., Cornblatt BA., Olvet DM., Mathalon DH., McGlashan TH., Perkins DO., Belger A., Seidman LJ., Thermenos HW., Tsuang MT., van Erp TG., Walker EF., Hamann S1., Woods SW., Qiu M., Cannon TD.

[32] "Harmonization of multi-site diffusion tensor imaging data", Fortin, J.P., Parker, D., Tun, 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.

[33] "Harmonization of cortical thickness measurements across scanners and sites", Fort in JP., Cullen N., Sheline YI., Taylor WD., Aselcioglu I., Cook PA., Adams P., Cooper C., Fava M., McGrath PJ., McInnis M., Phillips ML., Trivedi MH., Weissman MM., Shinohara RT., 2018.

[34] "Adjusting batch effects in microarray expression data using empirical Bayes metho ds", W. E. Johnson, C. Li, and A. Rabinovic, 2007.

[35] “Detecting and harmonizing scanner differences in the ABCD study”, Dylan M. Nielson, Francisco Pereira, Charles Y. Zheng, Nino Migineishvili, John A. Lee1, Adam G. Thomas and Peter A. Bandettini, 2018.

[36] “Data Harmonization for a Molecularly Driven Health System”, Jerry Ssu-Hsien Lee, Warren Alden Kibbe, Robert Lee Grossman, 2018.

[37] “Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data”, Griffith L., van den Heuvel E., Fortier I., Hofer S1, Raina P1, Sohel N., Payette H., Wolfson C., Belleville S., 2013.

[38] “Disability-free life expectancy: a cross-national comparison of six longitudinal studies on aging. The CLESA project”, N. Minicuci, M. Noale, S. Pluijm, M.Zunzunegui, T. Blumstein, D. Deeg, C. Bardage, M. Jylhä , 2004.

[39] “Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data”, Burns R., Butterworth P., Kiely KM., Bielak A., Luszcz M., Mitchell P., Christensen H., Von Sanden C., Anstey KJ., 2011.

[40] https://portal.dementiasplatform.uk/.

[41] “Development and use of a flexible data harmonization platform to facilitate the harmonization of individual patient data for meta-analyses”, Joeri Kalter, Maike G. Sweegers, Irma M. verdonck-de Leeuw,Johannes brug and Laurien M. Buffart, 2019.

[42] “Is rigorous retrospective harmonization possible? Application of the DataSHaPER a pproach across 53 large studies”, Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R. P., Knoppers, B. M., Hudson, T. J., Burton, 2011.

[43] “Dementias Platform UK (DPUK) Data Portal - supporting multi-modal data analysis, data linkage and real-world outcomes”, Christopher Orton, John Gallacher, Ronan Lyons, David Ford, Simon Thompson, Clare Mackay, et al., 2018.

[44] “Study protocol: the Whitehall II imaging sub-study”, Nicola Filippini, Enikő Zsoldo s, Rita Haapakoski, Claire E Sexton, Abda Mahmood, Charlotte L Allan, Anya Topiwala, Vyara Valkanova, Eric J Brunner, Martin J Shipley, Edward

Auerbach, Steen Moeller, Kâmil Uğurbil, Junqian Xu, Essa Yacoub, Jesper Andersson, Janine Bijsterbosch, Stuart Clare, Ludovica Griffanti, Aaron T Hess, Mark Jenkinson, Karla L Miller, Gholamreza Salimi-Khorshidi, Stamatios N Sotiropoulos, Natalie L Voets, Stephen M Smith, John R Geddes, Archana Singh-Manoux, Clare E Mackay, Mika Kivimäki & Klaus P Ebmeier, 2014.

[45] "Cohort Profile: The Whitehall II Study", Michael G Marmot, Eric Brunner, 2004.

[46] "Biobank: An Open Access resource for identifying the causes of a wide range of complex diseases of middle and old age", Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, Rory Collins, 2015.

[47] "Multimodal population brain imaging in the UK Biobank prospective epidemiological study", Miller KL., Alfaro-Almagro F., Bangerter NK., Thomas DL., Yacoub E., Xu J., Bartsch A., Jbabdi S., Sotiropoulos SN., Andersson JL., Griffanti L., Douaud G., Okell TW., Weale P., Dragonu I., Garratt S., Hudson S., Collins R., Jenkinson M., Matthews P., Smith SM., 2016.

[48] "The UK Biobank resource with deep phenotyping and genomic data", Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly & Jonathan, 2018.

[49] "Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank", Fidel Alfaro-Almagro, Mark Jenkinson, Neal K. Bangerter, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N. Sotiropoulos, Saâd Jbabdi, Moisés Hernández-Fernández, Emmanuel Vallée, Diego Vidaurre, Matthew A. Webster, Paul McCarthy, Chris Rorden, Alessandro Daducci, Daniel C. Alexander, Honghui Zhang, Iulius Dragonu, Paul M. Matthews, Karla L. Miller, Stephen M. Smith, 2018.

[50] "Predictive Modelling Analytics through Data Mining", Lakshay Swani, Prakita Tyagi, 2017.

[51] “Introduction to semi supervised learning”, Xiaojin Zhu and Andrew B. Goldberg, 2009.

[52] “Supervised and Unsupervised Machine Learning Algorithms”, Jason Brownlee, 2016.

[53] “Generalized Linear Models for Insurance Rating”, Mark Goldburd, Anand Khare, and Dan Tevet, 2016.

[54] “Ridge Regression for Solving the Multicollinearity Problem: Review of Methods and Models”, Hanan Duzan, Nurul Shariff, 2015.

[55] https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net.

[56] “Penalized Regression Essentials: Ridge, Lasso & Elastic Net”, Alboukadel Kassambara, 2018.

[57] https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net.

[58] “A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions”, Eric Schulz, Maarten Speekenbrink, Andreas Krause, 2017.

[59] https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319.

[60] https://scikit-learn.org/0.17/modules/gaussian_process.html.

[61] “Evaluating Machine Learning Models”, Alice Zheng, 2015.

[62] https://www.saedsayad.com/model_evaluation.htm.

[63] https://socialresearchmethods.net/kb/statinf.php.

[64] https://www.investopedia.com/terms/t/t-test.asp.

[65] “T test as a parametric statistic”, Tae Kyun Kim, 2015.

[66] “A guide to appropriate use of Correlation coefficient in medical research”, MM Mukaka, 2012.

[67] https://indatalabs.com/blog/predictive-models-performance evaluationimportant?cli_action=1572796194.637.

[68] https://git.fmrib.ox.ac.uk/fsl/funpack/.

[69] https://issues.dpuk.org/eugeneduff/wmh_harmonisation/tree/master/funpack_wmh_bb.

[70] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide.

[71] "A global optimisation method for robust affine registration of brain images", M. Jenkinson and S.M. Smith, 2001.

[72] "Improved optimisation for the robust and accurate linear registration and motion correction of brain images.", M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith, 2002.

[73] https://git.fmrib.ox.ac.uk/vaanathi/LOCATE-BIANCA/blob/master/LOCATE_User_Manual_V1.1_20052018.pdf.

[74] "Gaussian Processes for Big Data", James Hensman, Nicolo Fusi, Neil D. Lawrence, 2013.

[75] "An Introduction to Statistical Learning", G. James, D. Witten, T. Hastie and R. Tibshirani, 2013.

[76] "Spearman Correlation Coefficients, Differences between", Leann Myers and Maria J. Sirois, 2006.

[77] "High blood pressure, physical and cognitive function, and risk of stroke in the oldest old: the Leiden 85-Plus Study", Sabayan B, van Vliet P, de Ruijter W, Gussekloo J, de Craen AJ, Westendorp RG.

[78] "Longitudinal study of blood pressure and white matter hyperintensities: the EVA MRI Cohort", Dufouil C, De Kersaint-Gilly A, Besancon V, Levy C, Auffray E, Brunnereau L, Alperovitch A, Tzourio C.

[79] "Effects of systolic blood pressure on white-matter integrity in young adults in the Framingham Heart Study: a cross-sectional study", Maillard P, Seshadri S, Beiser A, Himali JJ, Au R, Fletcher E, Carmichael O, Wolf PA, DeCarli C.

[80] "Harmonising white matter hyperintensities measures across studies: impact of BIANCA training options", L. Griffanti, I. Bertani, V. Bordin, I. Mattioli, G.

Zamboni, S.Suri, E. Zsoldos, K.P. Ebmeier, M.M. Laganà, G. Baselli, M. Jenkinson, C.E. Mackay, E. Duff.

[81] “Between- and within-rater agreement in white matter hyperintensity segmentation from manual rating and a supervised automated classifier, FSL-BIANCA”, L. Griffanti, I. Mattioli, V. Bordin, I. Bertani, S. Suri, E. Zsoldos, K.P. Ebmeier, C.E. Mackay, G. Zamboni.