University of Nottingham
UK | CHINA | MALAYSIA

On harmonisation of brain MRI data across scanners and sites

Asante Ntata

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

December 2022

Abstract

Magnetic resonance imaging (MRI) of the brain has revolutionised neuroscience by opening unique opportunities for studying unknown aspects of brain organisation, function and pathology-induced dysfunction. Despite the huge potential, MRI measures can be limited in their consistency, reproducibility and accuracy which subsequently restricts their quantifiability. Nuisance non-biological factors, such as hardware, software, calibration differences between scanners and post-processing options can contribute or drive trends in neuroimaging features to an extent that interferes with biological variability and obstructs scientific explorations and clinical applications. Such lack of consistency, or *harmonisation* across neuroimaging datasets poses a great challenge for our capabilities in quantitative MRI. This thesis contributes to better understanding and addressing it. We specifically build a new resource for comprehensively mapping the extent of the problem and objectively evaluating neuroimaging harmonisation approaches. We use a travelling heads paradigm consisting of multimodal MRI data of 10 travelling subjects, each scanned at 5 different sites on 6 different 3.0T scanners from all the 3 major vendors and using 5 imaging modalities. We use this dataset to explore the between-scanner variability of hundreds of imaging-extracted features and compare these to within-scanner (within-subject) variability and biological (between-subject) variability. We identify subsets of features that are/are not reliable across scanners and use our resource as a testbed to enable new investigations which until now have been relatively unexplored. Specifically, we identify optimal pipeline processing steps that minimise between-scanner variability in extracted features (*implicit harmonisation*). We also test the performance of post-processing harmonisation tools (*explicit harmonisation*) and specifically check their efficiency in reducing between-scanner variability against baseline gold standards provided by our data. Our explorations allow us to come up with good practice suggestions on processing steps and sets of features where results are more consistent and reproducible and also set references for future studies in this field.

Acknowledgements

This PhD was funded by the Engineering & Physical Sciences Research Council. I would like to thank the directors of the Oxford-Nottingham Biomedical Imaging (ONBI) CDT program Prof. Peter Jezzard and Prof. Penny Gowland for delivering an excellent course and for their continual support.

I would like to express my deepest appreciation to my supervisor Prof. Stam Sotiropoulos for supporting me and guiding me each and every step of the way. I am grateful for his patience and for the role he has played in my development as an independent researcher.

I'm extremely grateful to Dr Olivier Mougin and Prof. Paul Morgan for the time they sacrificed in helping to acquire the MRI data and develop the scanning protocols for this project. I'd also like to thank research radiographers Jon Campbell and Andrew Cooper for their help in acquiring the data. I would like to extend my sincere thanks to the subjects who volunteered large portions of their time to be scanned in each of the scanners involved in this work.

Special thanks to Prof. Susan Francis and Dr Ludovica Griffanti for examining me and for their helpful feedback.

I'd like to thank God for never failing me and I'd like to recognise The Nottingham Bulwell Church for their encouragement and prayers.

Most of all, I would like to thank my family, especially my parents Harry and Joyce, my sister Faith, my aunt Carol and my cousins Delight and Divine for their belief in me and for being a continual source of support.

This thesis is dedicated to the loving memory of my cousin Freda.
You are forever in our hearts.

Contents

Abstract i

Acknowledgements ii

1 Introduction 1

1.1 Organisation of Thesis 5

2 Background 8

2.1 The Challenge of Quantitative Neuroimaging 8

2.2 Harmonisation Approaches 13

2.2.1 Harmonising Imaging Protocols 16

2.2.2 Harmonising the Raw Signal 20

2.2.3 Harmonisation of Imaging-derived Features 26

2.2.4 Summary of Existing Methods 36

2.3 Thesis Aims 37

2.3.1 Overcoming the Challenge of Evaluating Harmonisation Approaches 37

2.3.2 Building a Comprehensive Travelling Heads Dataset . . . 39

2.3.3 Mapping the Need for Harmonisation for Thousands of Multi-modal Imaging-derived Features 41

2.3.4 Testing Harmonisation Approaches 42

3 A Multi-modal Travelling-heads Harmonisation Resource for Brain MRI 43

3.1 Personal contributions 44

3.2 Introduction 44

3.3 Methods 46

3.3.1 Acquisition Strategy 46

3.3.2 Participants and Ethical Approvals 47
3.3.3 Overview of Multi-modal Acquisition Protocols 49
3.3.4 Acquisition Modifications to Accommodate all Scanners . 51
3.3.5 Total Readout Time for EPI 61
3.3.6 Data Quality Control 62
3.3.7 IQM Comparisons within scanners to Assess Consistency of scan-rescan 66
3.3.8 IQM Comparisons Between Scanners to Assess Consistency of Image Quality 66
3.4 Results 67
3.4.1 Examples of Collected Data 67
3.4.2 QC and Data Quality Comparisons 70
3.4.3 QC Across Scanners and Subjects 73
3.4.4 Within-vendor QC Differences 75
3.5 Discussion 79
3.6 Summary 81
3.7 Appendix: Modifications to Philips Achieva dMRI Data 82

4 Mapping Inter-scanner Variability for Multi-modal Imaging-derived Features 84

4.1 Introduction 85
4.2 Theory 88
4.2.1 UK Biobank Processing Pipeline and IDPs 88
4.2.2 T1-weighted Processing 90
4.2.3 T2-weighted Processing 91
4.2.4 SWI Processing 92
4.2.5 dMRI Processing 94
4.2.6 rfMRI Processing 97
4.3 Methods 99
4.3.1 Alterations Made to the UK Biobank Pipeline 99
4.3.2 Assessing IDP Cross-session Similarity 102

4.3.3 Testing for Scanner Effects Across IDPs 102
4.3.4 Mapping IDP Between-scanner Variability 103
4.3.5 Assessing the Preservation of Subject Ranking between-scanners 104
4.4 Results 105
4.4.1 Cross-session IDP Similarity 105
4.4.2 ANOVA Results 106
4.4.3 Mapping IDP Variability 109
4.4.4 Between-subject Ranking Consistency Across Scanners . 118
4.5 Discussion 121
4.6 Appendix: Bias in Ingenia dMRI Sessions 124
5 A Testbed for Evaluating Harmonisation Approaches 128
5.1 Introduction 128
5.2 Methods 132
5.2.1 Implicit Harmonisation 132
5.2.2 Explicit Harmonisation 137
5.3 Results 139
5.3.1 Implicit Harmonisation 140
5.3.2 Explicit Harmonisation 148
5.4 Discussion 150
6 Summary 158
6.1 Recommendations and Guidelines 158
6.2 Conclusions 160
6.3 Future Perspectives 162
Bibliography 165

Chapter 1

Introduction

The development of modern neuroimaging has given us novel insights into human brain anatomy, architecture and function for health and disease. Previously, such insight was mainly possible through labour-intensive microscopy techniques (e.g the use of sliced post-mortem tissue (Brodmann 1909)). While this method aided the understanding of brain structure and cytoarchitecture greatly, it was limited by its invasive nature and the tendency for tissue to be disrupted during extraction and preparation. In addition, the destructive nature of these measurements meant that it was not possible to monitor changes in-vivo, for instance imaging function and response to external stimuli, imaging longitudinal development, ageing or disease.

Magnetic Resonance Imaging (MRI) (Lauterbur 1973, Mansfield & Maudsley 1977) of the brain has made breakthroughs in addressing these limitations as it allows us to study the living brain non-invasively and in-vivo in healthy participants or patients. The technique offers great flexibility by allowing imaging contrasts that probe different anatomical, functional and physiological properties and mechanisms of neural tissue.

A common feature of all MRI modalities and contrasts is that they are mostly indirect. (See Figure 1.1). The signal that is measured is typically a proxy for

the quantity of interest and some processing or indirect inference may be typically needed to map what is measured to what one is interested in. This process can introduce challenges in quantifiability and consistency of MRI-derived measures (Schilling et al. 2021), such as the introduction of nuisance/uninteresting confounds, dependence on scanner hardware and software, and dependence on acquisition and processing steps that can reduce accuracy and precision of the measurements. Even fully anatomical images that aim to measure allometric changes can be confounded by MRI geometric distortions (Chang & Fitzpatrick 1990) that can vary across MRI scanners in non-trivial ways.

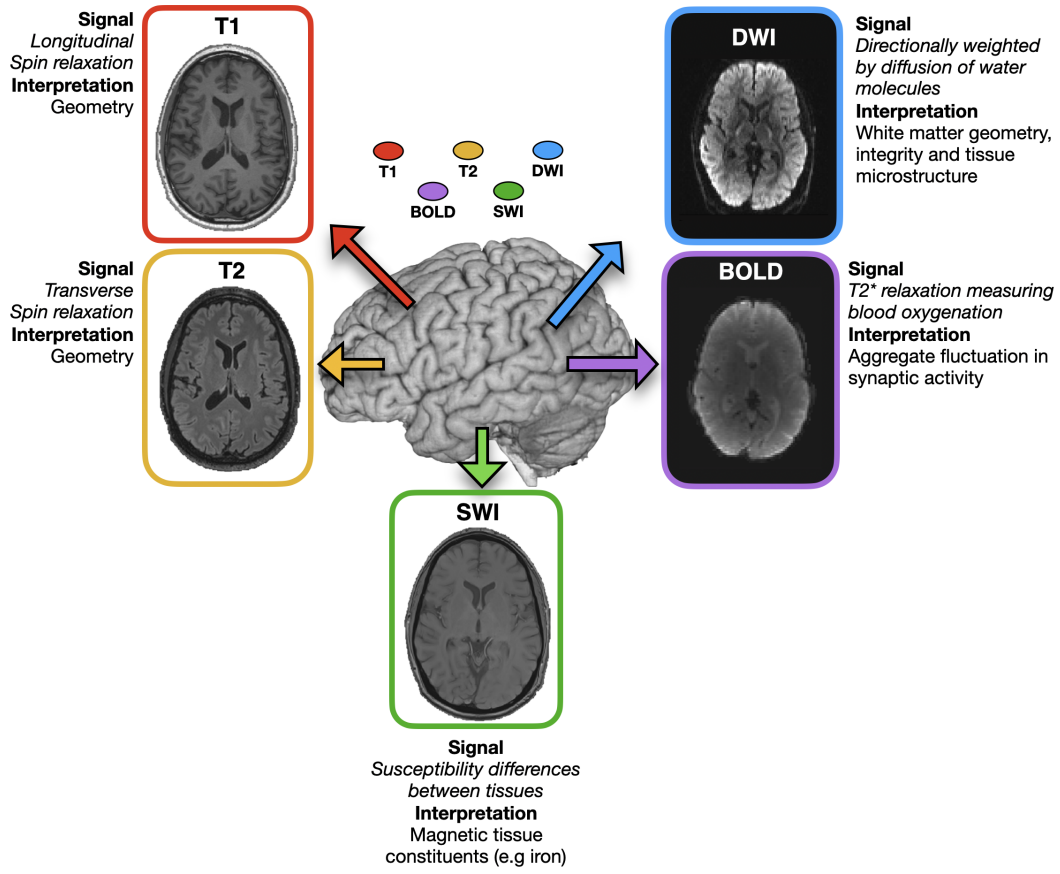

Figure 1.1: *Examples of the variety of contrasts that can be achieved and features that can be probed with various MRI modalities.*

These challenges in quantifiability can reduce the potential of MRI applications in a number of ways. Firstly, it complicates pooling data together from multiple studies and sites. The relatively recent emergence of the big data technologies and open science era has given rise to an increased number of

studies aiming to integrate numerous datasets across multiple sites, scanners and populations such as the UK Biobank (Sudlow et al. 2015), ADNI (Jack Jr et al. 2008), ABIDE (Di Martino et al. 2014) and ABCD (Casey et al. 2018).

The benefits of such multi-site collaborations for collecting neuroimaging datasets have been well reviewed (Van Horn & Toga 2009). These include increased access to different patient types and symptoms resulting in increased statistical power. Furthermore, multi-site collaborations are of great aid because the increased prevalence of machine learning in MRI (Davatzikos 2019) necessitates the availability of data sets which are sufficiently rich to train complex models on.

Secondly, reduced quantifiability can have downstream effects to reproducibility and accuracy of findings. Even in cases where the same dataset is being used, differences in analyses between teams can yield inconsistent results (Griffanti, Rolinski, Szewczyk-Krolikowski, Menke, Filippini, Zamboni, Jenkinson, Hu & Mackay 2016, Botvinik-Nezer et al. 2020). This issue is exacerbated when data are acquired or pooled from multiple sites as there is often a lack of reproducibility and consistency in quantitative MRI measurements from data (Zhu et al. 2011).

Thirdly, the lack of consistency or *harmonisation* across sites and scanners impedes MRI to be used quantitatively in clinical applications, for accurate patient-specific diagnosis and treatment monitoring. For a significant number of tasks, visual (and therefore subjective) inspection by the local radiologists is still the preferred way forward (Bruno et al. 2015). These problems can be alleviated by harmonising the data so that nuisance effects from scanner variability are removed/standardised; and imaging features become more consistent, reproducible and reliable. Developing harmonisation resources for multimodal brain MRI data across scanners and sites will be the subject of this thesis.

Harmonisation is a term that has been used to collectively describe approaches (including image acquisition and post-processing) that aim to remove unwanted inter-site/inter-scanner variability from data whilst preserving biological variability between subjects. This unwanted variability is caused by a number of nuisance factors such as scanning protocol, hardware and software, and data processing and can influence MRI measurements in non-predictable ways (Takao et al. 2011, Zhu et al. 2011). As shown with examples in Figure 1.2, such lack of harmonisation can cause bias and variance changes in ways that interfere with biological variability.

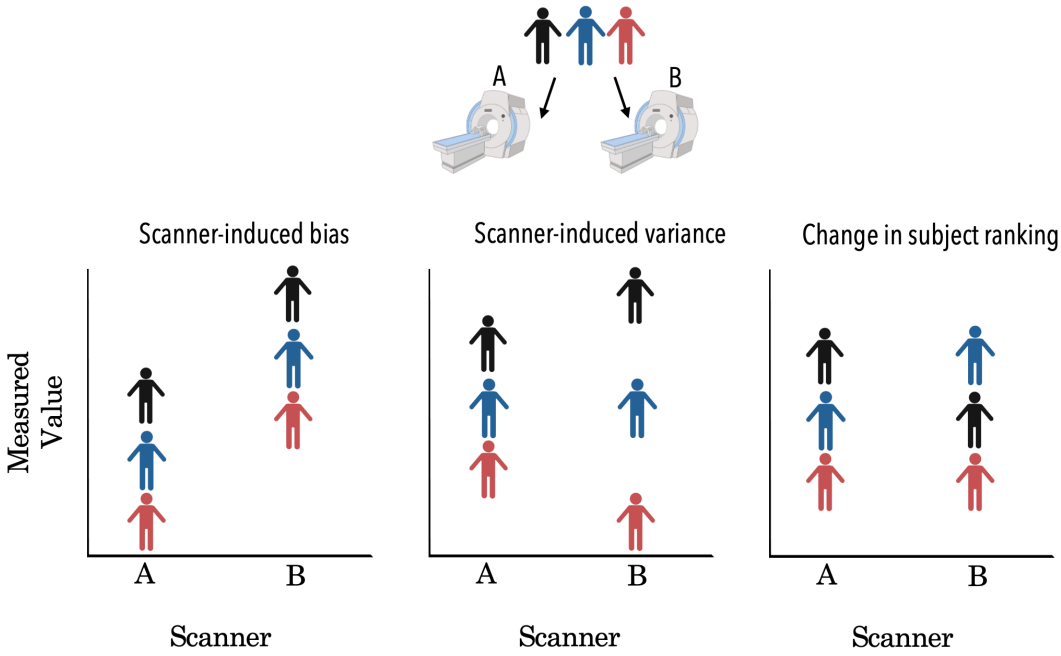

Figure 1.2: *Illustration of the potential challenges caused by a lack of harmonisation on imaging-derived features, such as scanner-induced bias, scanner-induced variance and change in subject ranking. The Figure shows an example of 3 different subjects each scanned on two different scanners.*

At the extreme end of these challenges, variability of measures obtained from the same subject but on different scanners can be as large as biological between-subject variability. For instance, the variability in diffusion MRI measurements between sites has been found to vary as much as 15% on the same subject, while effects of interest can be in the order of 5% (Mirzaalian et al. 2016).

In such extreme scenarios, imaging features obtained from the same subject being scanned in different scanners may look as though they are coming from different subjects, creating obvious interpretation issues and questions on usefulness of some of these metrics in real-world scenarios (Rao et al. 2017). Data which is so strongly associated with the site at which it was acquired is problematic because it becomes unclear whether observed variance in subjects is truly biological or merely a scanner related feature.

The goal of this thesis is to build a comprehensive resource for multi-modal MRI harmonisation and use it to provide novel insights and solutions to the relevant challenges. To do so we will use a travelling-heads paradigm, healthy individuals scanned multiple times across multiple imaging sites and MRI scanners. Compared to previous similar approaches this study is unique in considering scanners from all 3 major vendors, multiple generations of scanners within each vendor, within-scanner repeats, multiple neuroimaging modalities and 5 imaging sites in total. This resource will enable us for the first time to comprehensively evaluate existing processing options and harmonisation algorithms, map consistency and need for harmonisation across thousands of MRI-derived features and provide a testbed for new developments on reducing the problem.

1.1 Organisation of Thesis

This thesis is organised in 5 chapters. The next chapter gives a broad overview of literature on existing harmonisation approaches and groups them based on the commonalities between them. The following two chapters present the harmonisation resource we have built and how this can be used to map inter-site and inter-scanner effects for hundreds of imaging-derived, multi-modal features. The last research chapter presents an evaluation of existing harmonisation approaches and thus demonstrating the power of the harmonisation

resource, as well as novel insights into harmonisation and processing performance. A final chapter summarises the results.

Specifically, **Chapter 2** is a background chapter that reviews the current literature and the harmonisation approaches that have been proposed. The approaches are split into three groups: 1) techniques which harmonise imaging protocols, 2) techniques which remove scanner specific differences from the raw data and 3) techniques which harmonise imaging-derived features.

Chapter 3 is the first original research chapter and presents the non-trivial setup for data collection using a travelling-heads paradigm and 6 scanners in total, as well as the acquisition protocols and quality control process. It gives an overview of the scanners and sites used as well as the number of subjects for which different data were acquired. An overview of the imaging modalities is also given including protocol details and sequence information. Quality control (QC) comparisons across the scanners and modalities are also presented with explorations on how quality differs between vendors due to specific hardware features.

Chapter 4 builds upon the travelling-heads datasets and explores how inter-site and inter-scanner effects influence hundreds of multi-modal imaging-derived features. An overview of the image processing pipelines and imaging features derived are presented. The between-scanner variability of various imaging features is explored and this is compared to within-scanner variability and biological (between-subject) variability. The effects of using different vendors is demonstrated through various comparisons including the effect on between-scanner ranking variability. This chapter demonstrates and summarises modalities and groups of imaging-derived features that seem to be more and less robust against inter-site/inter-scanner effects.

Chapter 5 is the last research chapter and demonstrates the value of the built resource when used as a testbed to objectively evaluate for the first time existing harmonisation approaches. One way is by identifying the optimal pipelines and processing steps that minimise between-scanner variability in extracted features compared to e.g. within-scanner variability or biological variability (implicit harmonisation). Another way is testing performance of post-processing harmonisation tools (explicit harmonisation) and specifically checking whether the harmonised features between-scanners are indeed less variable (and by how much) compared to no harmonisation. This chapter allow us to come up with good practice suggestions on processing steps and sets of features where challenges are mitigated.

Finally, **Chapter 6** summarises the thesis and discusses potential directions for the future.

Chapter 2

Background

2.1 The Challenge of Quantitative Neuroimaging

Imaging the brain using magnetic resonance imaging (MRI) offers great potential and flexibility by allowing imaging contrasts that probe different anatomical, functional and physiological properties and mechanisms of neural tissue. However, measurements are indirect, reflecting directly properties of water molecules within the brain tissue probed in different ways. Hence a mapping of these measurements to biophysical properties of interest may be needed. Moreover, due to the spatial scale of imaging, images reflect macroscopic views of tissue and measurements frequently reflect thousands of underlying processes or microstructures co-occurring within the same observation window. This makes quantifiability of measurements challenging.

MRI measurements are inherently noisy. Thermal noise (Gudbjartsson & Patz 1995) induces scan-rescan variability in images of the same individual acquired multiple times from the same scanner (Wrobel et al. 2020). In addition to this *within-scanner* variability, the flexibility of MRI scanners can potentially add to these challenges. Particularly with modern acquisitions and scanning pro-

tocols, hardware and software differences between scanners of the same or different manufacturers can be significant and can add further *between-scanner* variability when scanning the same individual across multiple settings (Han et al. 2006, Zhu et al. 2011).

An attempt to mitigate for these effects is performing quality control and assessment of scanners using **physical phantoms**. A range of phantoms have been developed, both experimental (Palacios et al. 2017) and commercial (Laun et al. 2009). Phantoms have known geometry and structural properties and can be used to standardise (or ensure compliance with prior standards). For instance, they can be useful in ensuring geometric distortions in MRI are within acceptable limits (Chen et al. 2004), that the gradients are well calibrated (Bagherimofidi et al. 2019), or that biophysical properties like diffusion coefficients (Zhou et al. 2018) can be accurately mapped (Keenan et al. 2016). Phantoms however are limited to what level of variability they can capture and how realistic they can be. For example, phantoms can mimic only a certain type and range of biophysical properties due to construction complexity/infeasibility, they cannot easily capture dynamic biophysical properties or artefacts related to scanning living humans, such as physiological noise, subject motion and their interaction with other distortion fields or quantitative measurements.

As a consequence, variability in MRI measurements and difficulty to standardise remains a challenge. Measurements can be influenced by **"nuisance factors"** such as the scanning protocol, hardware and software which are different between vendors and can vary with site (Han et al. 2006, Zhu et al. 2011). This lack of *harmonisation* is not simply an inconvenience, but a limiting factor in many occasions for scientific and clinical applications. It has been shown that variability of measures obtained from the same subject but on different scanners can be as large as biological between-subject variability.

For instance, the variability in diffusion MRI measurements between sites has been found to vary as much as 15%, while effects of interest can be of the order of 5% (Mirzaalian et al. 2016).

This has been further supported by studies looking into predictability of scanning site from imaging data. When features are derived from images acquired from multiple scanners, they can show strong association with the scanner on which they were acquired, rather than with the subject being scanned. This was demonstrated in (Fortin et al. 2018) where linear discriminant analysis was used to find a linear combination of features which separate the data into two or more classes. The study showed that cortical thickness measures cluster perfectly by site rather than by subject. Further evidence that imaging data is strongly associated with the scanner on which it has been acquired has been found by playing the *"Name that dataset"* game (Torralba & Efros 2011). The goal of the game is to train a classifier to guess the dataset that an image has come from. A perfectly unbiased dataset should have a classification accuracy which is as good as random chance. This game was applied to neuroimaging data in (Wachinger et al. 2021) and results showed that brain scans from 17 large-scale public datasets could be correctly assigned to their respective datasets with 71.5% accuracy. Similar findings were found in (Glocker et al. 2019) where a random forest binary classifier was trained to distinguish the origin of T1-weighted images. The classifier was able to predict the origin of the data with a high degree of accuracy confirming the presence of site effects in the data.

A way of addressing the complications of having data from multiple scanners is to *attempt to keep the same scanner* model/vendor throughout a study. However it has been shown that even with scanners of the exact same model, variability in imaging features can exist (Takao et al. 2011). Figure 2.1 has been reproduced from (Wrobel et al. 2020) and visually illustrates how much

variability can exist even in such a controlled setting. The histograms of voxel intensities for 7x2 (scan and rescan) data on single subject across 7 sites is shown. We see that not only are differences observed between different scanners but also within the same scanner. To the extreme of this approach, one could attempt to use a single scanner throughout, which was the approach originally taken by the UK Biobank (Miller et al. 2016). This may work as a bespoke setting, but it is not a sustainable way forward.

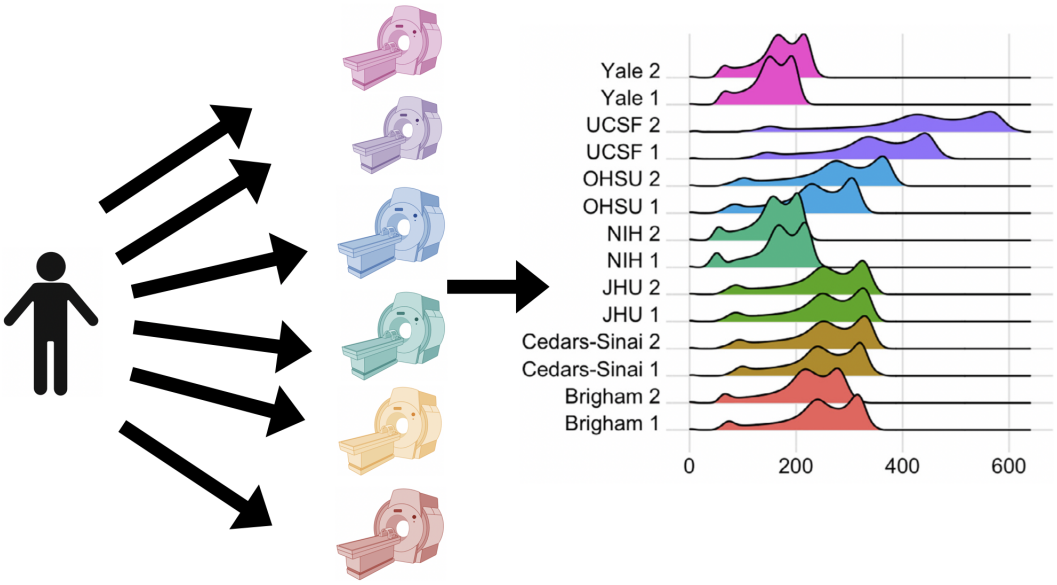

Figure 2.1: *Histograms of voxel intensities for scan-rescan data on a single subject across 7 sites. For each site, a repeat scan has been acquired and the 14 histograms show the 7x2 (scan and rescan) intensity distributions for a T1-weighted acquisition. Figure reproduced from (Wrobel et al. 2020)*

.

The challenges caused by lack of harmonisation in acquired image *intensities* (as seen in the previous figure for instance) can be further exacerbated when considering imaging-derived *features*. Due to the indirect nature of MRI measurements, a large number of steps may be needed to map intensities to features (for instance BOLD signal to neuronal activation summaries in functional MRI, diffusion-sensitised signal to tracts in diffusion MRI, grey-to-white-matter contrast to cortical thickness in anatomical MRI). A number of processing tools and pipelines exist (Smith et al. 2004, Glasser et al. 2013, Alfaro-Almagro

et al. 2018, Friston 2007, Esteban, Markiewicz, Blair, Moodie, Isik, Erramuzpe, Kent, Goncalves, DuPre, Snyder et al. 2019, Desikan et al. 2006, Fischl et al. 2004) as this mapping is not straightforward and typically needs multiple steps of processing and modelling. Variability in the signal is then propagated down the processing routines to variability in the features. However the tools themselves can add a second layer of variability. It has been well reported that even in cases where where the same dataset is being used (Griffanti, Rolinski, Szewczyk-Krolikowski, Menke, Filippini, Zamboni, Jenkinson, Hu & Mackay 2016, Botvinik-Nezer et al. 2020, Schilling et al. 2021), differences in analyses, processing tools and pipelines can yield inconsistent results. The combination of pooling data from multiple sites in conjunction with the different permutations of available processing steps are potential contributors in what has been termed a "reproducibility crisis". It is therefore important to concurrently consider both steps (acquisition and processing) when attempting to solve the harmonisation problem.

In summary, lack of harmonisation in imaging protocols and features is a significant bottleneck for realising the full potential of brain MRI for a number of applications that rely on quantifiability of imaging measurements. In science it can lead to lack or difficulty in reproducing studies and results. Lack of harmonisation is also a significant obstacle for pooling together the plethora of MRI data that are available across imaging facilities and using them in modern deep learning applications. In the clinic, it impedes MRI to be used quantitatively for accurate patient-specific diagnosis and treatment monitoring. For a significant number of tasks, visual (and therefore subjective) inspection by the local radiologists is still the preferred way forward. Therefore, providing frameworks and data that subsequently allow to map, evaluate and solve the problem would be a significant contribution and will be the subject of this thesis.

2.2 Harmonisation Approaches

Over the years, a number of harmonisation approaches have been developed to explicitly mitigate for or reduce these inter-site/scanner effects. In this section we overview representative groups of these approaches. Collectively, these can be divided into three groups:

a) Techniques which explicitly aim to harmonise imaging protocols.

b) Techniques which aim to remove scanner specific differences from the raw signal using post-processing.

c) Techniques which aim to remove the scanner differences from derived imaging-derived features using post-processing.

The first group of methods aim to match as faithfully as possible the imaging protocols. Even if in principle this is possible, certain aspects of modern acquisitions are a priori difficult to match. For instance in-plane acceleration reconstruction algorithms can have many different implementations across vendors, or different types of filters are adopted by different manufacturers and these implementations affect and change in different ways the statistical properties of the signal (Dietrich et al. 2007*a*). This makes perfect matching impossible. For this reason, post-processing techniques have been proposed that attempt to remove or reduce inter-scanner effects post-acquisition. These operate either at the raw signal level or at the derived feature level.

Within these 3 broad categories there are various sub-categories which contain their own harmonisation approaches. For example, approaches which aim to harmonise the raw signal can be divided into intensity normalisation, alternative representations of data (such as spherical harmonics) etc. An overview of this structure is shown in Figure 2.2. The following sections expand on this structure and provide more detailed information on each of the approaches.

An alternative way of dividing harmonisation approaches which will be used extensively in Chapter 5 is by categorising them either as *implicit* or *explicit* harmonisation. *Explicit* harmonisation refers to post-processing harmonisation tools which aim to remove or mitigate scanner/site effects. Groups **b** and **c** from the paragraph above fall within this category. *Implicit* harmonisation refers to applying optimal pipeline processing steps which minimise between-scanner variability in extracted features. These approaches are not covered in this chapter as they are not harmonisation tools in their own right but rather constitute a framework for the selection of tools.

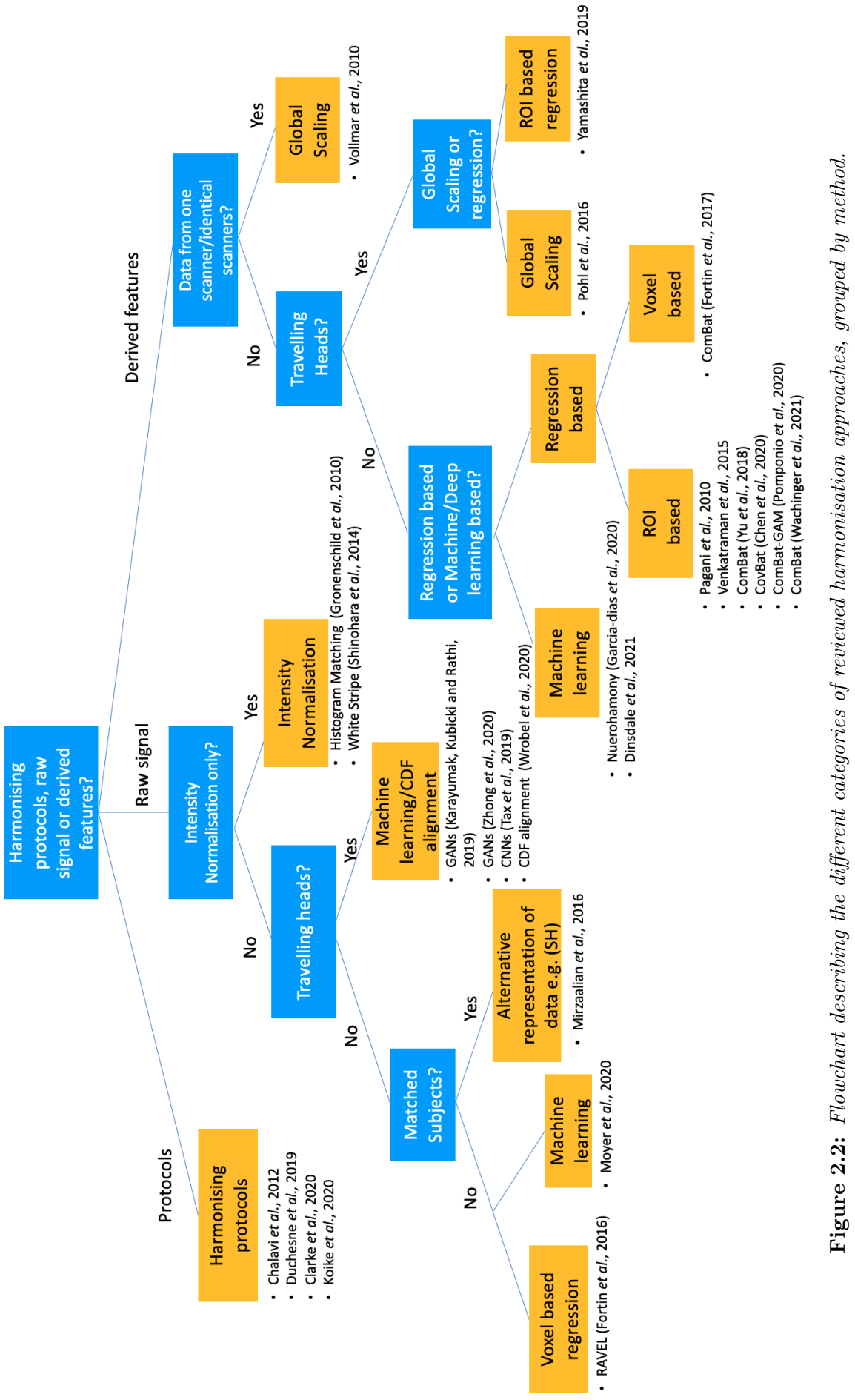

Figure 2.2: *Flowchart describing the different categories of reviewed harmonisation approaches, grouped by method.*

2.2.1 Harmonising Imaging Protocols

A major cause of unwanted variability in MRI acquisitions is poor matching between contrast sensitive parameters such as echo times, repetition times, flip angles and inversion times. Contrast is a key factor in determining the outcome of brain analysis tools such as segmentation and registration, so it is important to reduce, as much as possible, differences in these parameters. When harmonising neuroimaging protocols, acquisition parameters are carefully chosen so that any interaction between effects of interest and sequence parameters is minimised. This is demonstrated by (Chalavi et al. 2012) in a 2-site study for T1-weighted volume protocols. A range of pulse sequence parameters were assessed, based on image quality and the reproducibility of cortical and subcortical volume measurements, until the optimum parameters were identified.

In (Duchesne et al. 2019), protocols for T1-weighted and T2-weighted scans were harmonised across a 16 different scanners spanning the 3 major vendors (Siemens, Philips, GE). Protocol parameters such as echo time and repetition time were chosen to obtain images of similar quality in terms of contrast and resolution. Despite this focused effort, inherent differences in implementations and hardware between vendors resulted in significant differences in T1-weighted SNR values and cortical and sub-cortical CNR values. However, no significant differences were found in total brain volume measures suggesting that the harmonisation of protocols was effective, at least for these very specific features.

More recently, a substantial effort has been made to harmonise multi-modal protocols at 3.0T, including T1- and T2-weighted, resting-state functional and diffusion-weighted MRI (Koike et al. 2021). Travelling heads (healthy subjects scanned in multiple scanners) data was acquired in 5 sites and results

showed comparable results for three imaging-derived features: cortical thickness, myelin (T1-weighted/T2w ratio) and functional connectivity. Even if the study used only scanners from the same vendor (Siemens), yet there was high between-scanner variability for the more complex features (myelin and functional connectivity) as shown by Figure 2.3 which shows the correlation matrices of the parcellated cortical thickness and myelin (not biasfield corrected [non BC]) four travelling subjects, scanned by five scanners/sites. It can be seen that the same subject for more advanced metrics (myelin) can be as similar to other subjects as to itself, even when the same vendor scanners have been used

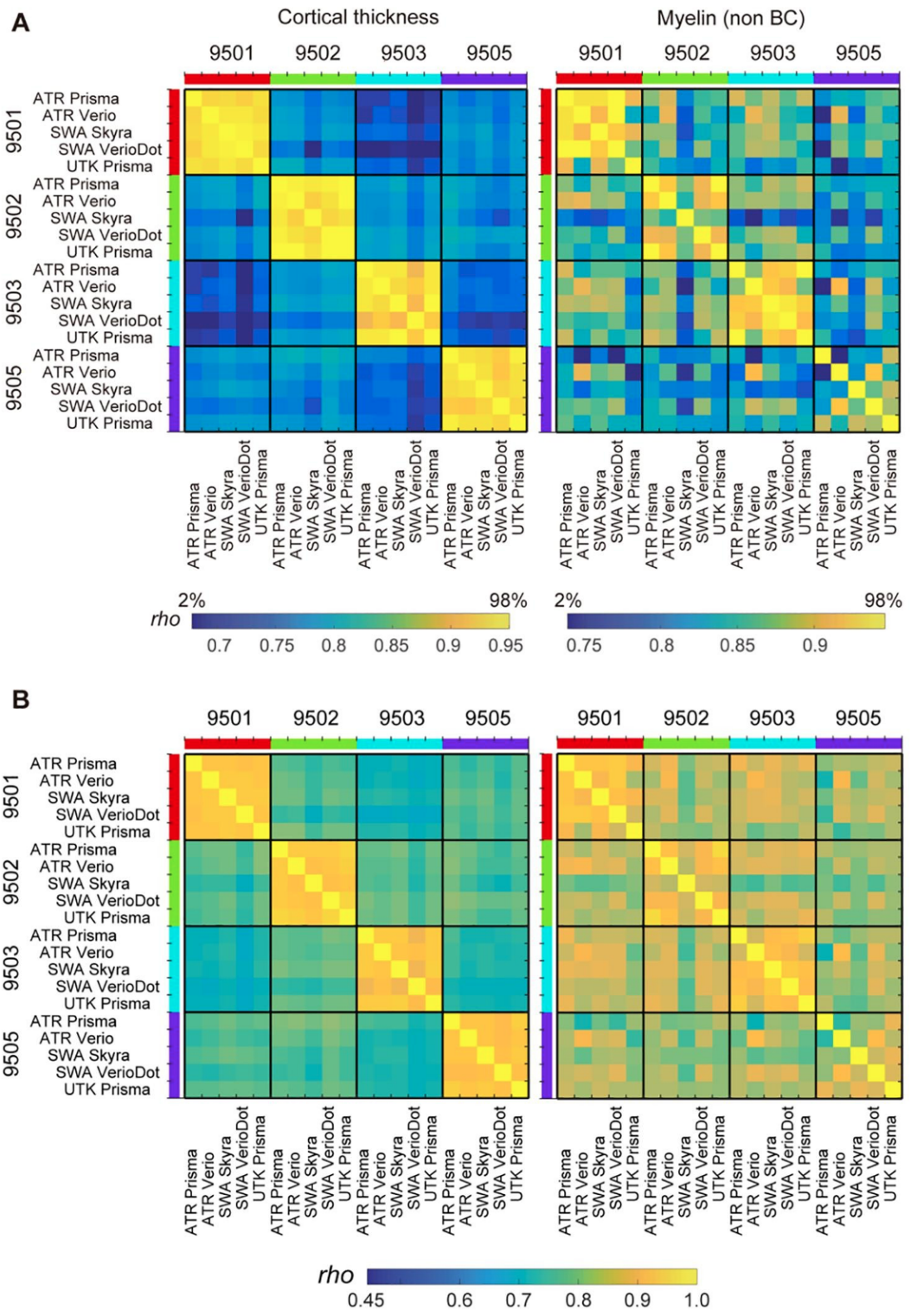

Figure 2.3: *Correlation matrices of parcellated cortical thickness and myelin of 4 travelling subjects, scanned in 5 scanners/sites. Figure shows the difference in consistency between cortical thickness and myelin. A) Colour ranges are scaled by the distribution of the correlation coefficients (2% to 98% of histogram) to highlight the contrast between 'within-subject' similarities and 'across-subject' similarities, while in the lower row (B) Colour ranges are scaled by the same absolute values across all modalities. Spearman's correlation coefficient (rho) is shown using a colour bar placed at the bottom. Non BC: non biasfield corrected. Figure from (Koike et al. 2021).*

A similar procedure was carried out for higher field scanners by the UK7T Network to harmonise protocols across 5 different sites consisting of 7T scanners covering 3 different models provided by 2 different vendors (Clarke et al. 2020). Protocols were harmonised for T1-weighted, T2-weighted, T2*-weighted contrasts and resting-state functional MRI. Data from 1 subject scanned in each of the 5 sites was used to assess the quality of the harmonisation using various metrics. The protocol harmonisation successfully reduced inter-site variation of cortical thickness for T1-weighted images to below 5% in all cortical regions. Inter-site differences in susceptibility and for T2* derived measures were also reduced to low levels. For the task and resting-state fMRI, the harmonisation of the protocols across scanners was successful to the level of the within-scan variability of a single subject; meaning there was no significant increased variance in % BOLD measurements in selected ROI's taken across sites compared to repeats at a single site.

In summary, a number of coordinated efforts have shown potential and are in the right direction. The inherent difficulty however to exactly match different acquisition implementations across scanners and manufacturers makes this an unresolved problem in the general case. There is also the possibility of "nominal" rather than true matching of acquisition parameters, as manufacturers do not typically share exact implementation details of their product sequences, hence there is a source of uncertainty on what exactly each acquisition parameter reflects in each scanner. In addition, there is a lack of standardised ways to evaluate success for these harmonised protocols and for the thousands of features that can be potentially derived from brain images. Current efforts on harmonising protocols can be considered a good first step, but as not fully addressing the challenge, different types of post-processing approaches have been developed. We overview a representative set of those in the following sections.

2.2.2 Harmonising the Raw Signal

A first family of post-processing methods aim to directly harmonise the raw signal as a post-acquisition task. This can be seen as a pre-processing step aiming to remove or mitigate between-scanner effects, which can be linear or nonlinear. We overview some representative examples below.

2.2.2.1 Intensity Normalisation Methods

Histogram matching

One of the simplest approaches to harmonising raw signal involves normalisation of measured intensity values by matching histograms. The aim of this approach is to address the issue that image intensities of separately acquired images can vary even if they are images of the same patient, obtained on the same scanner, for the same body region and with the same protocol. (Nyúl & Udupa 1999) proposed an approach which involved histogram matching. By averaging the intensity histograms of a reference population a template histogram is created onto which the individual histograms of each subject can be mapped. This method was used by (Gronenschild et al. 2010) to standardise image intensities for segmentation of cortical structures from T1-weighted images acquired on the same scanner. Such methods are useful in some limited scenarios but violate "the principles of image normalisation" as defined in (Shinohara et al. 2014). A major limitation is that the success of the method relies on the assumption that the distribution of tissue type is the same across subjects and scanners meaning it will not generalise well to studies involving multiple scanners.

Alignment cumulative distribution functions

In (Wrobel et al. 2020), data from *travelling heads* were leveraged to align the cumulative distribution functions (CDF) of image intensities from images of the same subject acquired in different scanners to a template. This method

is based on the assumption that the variability in voxel intensities between scanners is driven by scanner effects rather than biological differences and that this difference can be learnt and removed by aligning the distribution of intensities of each subject to a subject-specific template (for each subject a baseline scan taken taken at one of the sites was used as the common template to which all CDFs within a modality were aligned) This alignment was performed via a warping function which is estimated using linear interpolation between the cumulative distribution functions of the subject and the template at equally spaced intensity values. The authors used this method in conjunction with White Stripe (Shinohara et al. 2014) (defined in the next section) and they compare it to the use of White Stripe alone. The efficacy of this method was assessed by comparing T2 lesion volumes for scan-rescan pairs at each of 7 scanners, comprising of Siemens systems. On average, this approach outperformed the use of White Stripe alone though in some individual cases white-stripe performed better.

Harmonisation by reference region: White Stripe

Expanding upon the idea of histogram matching, (Shinohara et al. 2014) proposed a method that minimises the discrepancy between the distribution of intensities, while being robust to the effects of acquiring data in multiple sites. The method, coined White Stripe, uses a section of normal appearing white matter as a reference tissue region of interest (ROI), to represent a region least affected by partial volume averaging. The properties of the distribution of the reference ROI are used to appropriately adjust the distributions in other tissue ROIs accordingly. The method was assessed on the intensities of WM and GM areas from T1-weighted and T2-weighted acquisitions. The results of this method demonstrated increased comparability of white matter histograms across subjects, but it showed sensitivity to the choice of the reference region, particularly for harmonising grey matter intensities.

2.2.2.2 Voxel-based Regression-based Approaches

RAVEL

In order to address the limitation of assuming constant correction factors within ROIs (as described above), (Fortin et al. 2016) proposed RAVEL, which takes into account scanner differences at the voxel level. This method is an extension of White Stripe (Shinohara et al. 2014) and it was tested on the intensities of CSF, WM and GM derived from T1-weighted images. The approach is to first define a set of control voxels, which are assumed to have relatively constant intensity. Any variability within these voxels is interpreted as reflecting unknown/unwanted site effects. In that study CSF voxels were chosen as control voxels because they are not associated with disease. Therefore, any variation in these is assumed to be non-biological. A regression model is then used at each voxel to regress out these reference intensities, hence removing variance explained by non-biological variation in these control voxels. (Fortin et al. 2016) applied RAVEL in conjunction with White Stripe, and found that after the correction, the replicability of the voxels known to be associated with pathology improved compared to just using White Stripe and histogram matching. This approach proved successful on structural T1-weighted images but it did not work well for other modalities, such as diffusion MRI. This was shown in (Fortin et al. 2017) where RAVEL (used because FA values in CSF should be near 0 for the participants of any study) was successful at reducing the variability in FA values but it was unable to account for local changes in MD values. The suggested reason for this was that there was a lack of correlation between average CSF values and MD values which prevents CSF intensities from being used as a suitable reference for standardisation of MD maps.

2.2.2.3 Alternative Signal Representations: Spherical Harmonics

The first non-scaling-based work that explicitly addressed harmonisation without the use of statistical covariates was performed by (Mirzaalian et al. 2016). The premise of the approach was to map the signal from different sites to a single site. During the mapping, one of the scanners was selected as the reference scanner and the scanner to be harmonised or mapped to the reference was selected as target scanner. The scanner differences were captured using rotational invariant spherical harmonic (RISH) features using spherical harmonic coefficients. These features represent the energy of the signal at different angular frequencies. The harmonisation is done by a two-step process: A learning part and a harmonisation part. The learning part uses age-and-sex-matched controls to create scale maps of the voxel-wise average RISH features of each scanner. The harmonisation step then applies these maps to the data from the target site to modify the signal. This method was performed on diffusion data and it was successfully validated on a travelling subject scanned on 6 different scanners with results indicating that statistical differences had been removed. The limit of this method is that it relies on age-and-sex-matched controls, rather than travelling subjects, to learn scanner differences which is not ideal since matching is not always feasible and the features to match can be very study-specific.

2.2.2.4 Machine/Deep Learning Methods using Travelling Subjects

In light of the limitations of using matched subjects, the ideal case is to have travelling subjects scanned at each site; i.e. the same volunteers that are scanned using different scanners, providing therefore measurements that reflect how the same individual is depicted by different machines. This allows for the use of advanced nonlinear strategies such as machine or deep learning to capture between-scanner differences. In a ***supervised learning*** paradigm, using the travelling subject data for training, the algorithms aim to synthesise

images from one scan to be similar to the images from a target scanner. The success of the methods can be assessed against the ground truth from the actually acquired data in the target scanner.

A deep learning approach using travelling subjects was proposed by (Karayumak, Kubicki & Rathi 2019) for diffusion data. The aim was to learn a nonlinear mapping of Prisma to Connectom scanner using paired RISH features. The nonlinear mapping between data of multiple scanners was learnt using a cycle generative adversarial network (CycleGAN) with Prisma RISH features as inputs and Connectom RISH features as outputs. GANs use two neural networks: a generator and a discriminator. The training is done so that the former generates an image that so closely resembles an image of interest with the aim of convincing the latter it is genuine, while the discriminator is trained to identify synthetic from real data. In this way, the generator eventually estimated an input-to-output scanner mapping. Once this mapping was learnt, it was applied to the scans of each subject. In the mentioned study, RISH features of 16 Human Connectome Project (Van Essen et al. 2013) healthy subjects with dMRI scans obtained from both Siemens Prisma and Skyra Connectom scanners were used to capture the scanner differences in the training phase and this was assessed on untouched data from 1 of the subjects in the testing phase. This was done 16 times leaving a different subject out each time. The results showed that existing scanner differences were minimised after harmonisation almost to the level of within-scanner variability. A limitation of the RISH features method, noted by (Zhong et al. 2020) is that the accuracy of the transformation from diffusion weighted images to the representation of a spherical harmonic basis depends on a the number of diffusion-sensitising gradient directions (Descoteaux et al. 2007). Such direction numbers may not always be achievable or feasible in some studies. Therefore, (Zhong et al. 2020) propose a method which uses GANs also but is applied on DTI derived metrics themselves rather than raw signal representations so that the need of a high

number of gradient directions is removed.

Other deep learning methods trained on data from travelling subjects are presented in (Tax et al. 2019) where algorithms were developed as part of a competition. Diffusion data from 14 subjects scanned in 3 different scanners were acquired. Most of the algorithms were based on convolutional neural networks (CNNs), which were trained on 10 of the subjects to learn the mapping of the data sets across scanners. The remaining 4 subjects were used to assess if this learnt mapping generalised to unseen data. These networks consist of layered structures where input data is propagated through intermediate layers before reaching an output. The goal of the network is to learn the input-to-output mapping to minimize error in the output. This learning happens during a training period where parameters in intermediate layers are adjusted so that output data for each training sample, well approximate the data from the actual acquired data. This mapping was then used to estimate what the data at a target site would look like when given data from a different site as an input. While the algorithms compared in this study managed to reduce multi-site variability, the authors pointed out that it would be important to perform more rigorous testing e.g. in other training sets. In addition, this study was performed on single-shell data and performance on multi-shell data is currently unknown. As well as this, the method relies on brains having been scanned at all sites. (Moyer et al. 2020) point out that while scanning subjects at multiple sites is advisable to validate harmonisation methods on such data, it is advisable to limit reliance on this due to the fact that data acquired in this way are usually not available (and therefore not representative of often encountered scenarios) and are costly to acquire.

To address the limited availability of travelling head data, (Moyer et al. 2020) proposed a deep learning method which follows an ***unsupervised learning*** paradigm in that it does not require multiple images of the same subject ac-

quired at different sites. Their approach was to map diffusion data from one scanner to another so that, given images from one site, it can be predicted what they would have looked like had they come from another. The role of the neural networks is to encode image data from a target to an intermediate representation which is invariant to any scanner. A separate neural network is then used to decode the information from the intermediate representation and reconstruct it to look like data from a site of choice. During training, parameters in the network were adjusted such that any reconstruction made from the intermediate site is still maximally relevant to the input data but contains no information about the site of origin. Therefore, a crucial component of the learning period was to make sure that any attempts at predicting which site the reconstructed data came from performed no better than random chance. This method was shown to outperform (Mirzaalian et al. 2016). A limitation however, is that it was demonstrated only in white matter. It's performance in grey matter is yet to be assessed.

2.2.3 Harmonisation of Imaging-derived Features

The above approaches focus on harmonising directly the image intensities or representations of them. Alternatively, a group of techniques aims to reduce between-scanner variability of imaging-derived features, such as subcortical structure volumes or microstructural measures in major white matter tracts. These features are usually obtained after the data have undergone a series of processing steps, such as distortion corrections, modelling and registration to an atlas. Techniques which harmonise these features fall into various subcategories, as we see below.

2.2.3.1 Global scaling approaches

One of the simplest approaches is to first investigate bias in derived features between the sites. If there is consistent bias, a scaling factor can be applied to each site to mitigate it. This was the method used by (Vollmar et al. 2010) where they corrected for consistent inter-site bias of fractional anisotropy measures in a 2-site study. The corrections they applied successfully shifted the individual distributions of values obtained for each site and therefore indirectly reduced the variance of inter-site feature values to the point that it was not significantly different to the variance of intra-site results. The limitation of this study was that it considered two "nominally identical" scanners with identical acquisition protocols which, in principle, guaranteed identical hardware, software and firmware between the two sites. Hence, it assumed that the variability of measurements in each site is identical, which may not be true in general. Furthermore, this approach assumes that a global linear scaling can explain all differences, but as pointed out in (Fortin et al. 2017), site differences are typically region-specific and global scaling approaches are insufficient to correct for such effects.

An alternative approach suggested by (Pohl et al. 2016) was to use "human phantoms" (i.e. travelling heads) to obtain an estimate of the scaling factor between scanners. Their analysis reported tract based spatial statistics of fractional anisotropy, mean diffusivity, axial diffusivity, and radial diffusivity. In total, 3 travelling heads made 26 visits across different sites. They then calculated a ratio of the mean values across the visits and used this as a linear correction factor which they applied to one of the scanners for each of their obtained features. This method was successful in reducing the standard deviation of DTI metrics for FA by almost half. Apart from the logistical challenge of acquiring scans from subjects across all scanners used in a study, a further limitation, as pointed out in (Mirzaalian et al. 2016) is that scanner related

effects can be nonlinear and their effects vary across regions. This method can therefore lead to erroneous results in aggregating data with nonlinear differences of which brain data is a typical example.

2.2.3.2 Regression-based approaches (voxel-based)

A major limitation of global scaling methods is that they depend on an idealised scenario where the scanners and acquisition protocols are identical. Regression based approaches attempt to overcome this assumption. This group of methods aims to fit regression models to feature values in ways which separate biological effects and site effects into different variables. The fitted regression aims to separate the site effects from the biological features of interest by considering the site effects as regressors in a model. This model is then used to calculate updated values which are free of site-effects. For instance, a linear regression model would take the form

$$Y_{ijf} = \alpha_f + \gamma_{if} + \epsilon_{ijf} \tag{2.1}$$

where Y_{ijf} is an image-derived measurement for imaging site i, subject j, and feature type f. α_f is the average value of the feature, γ_{if} is an additive site effect and ϵ_{ijf} is the variance. In this case, the parameters would be estimated by performing a least square regression and the harmonised value of the feature would be the residual:

$$e_{ijf} = Y_{ijf} - \hat{\gamma}_{if} \tag{2.2}$$

the input features could be, for example, an array of cortical thicknesses or fractional anisotropy values for a subject scanned in multiple scanners. This can be performed within regions of interest or on the voxel level.

ComBat

A major harmonisation approach in the field is ComBat (Fortin et al. 2017).

Originally used to adjust for batch effects in genomics data (Johnson et al. 2007), ComBat has been adapted to remove differences in data coming from different scanners. The method has also been shown to work in structural data for harmonising cortical thickness measurements (Fortin et al. 2018), diffusion derived measures (Fortin et al. 2017) and functional measures (Yu et al. 2018). ComBat models imaging feature measurements as a linear combination of the biological variables and the site effects. The model assumes that site effects exist when comparing across sites the mean and variance of (voxelwise or ROI) derived features. The "true/harmonised" mean and the variance of each feature are treated as parameters drawn from a prior distribution, and are assumed to be common across all sites when harmonised. This prior represents an initial estimate of the distribution of the mean and variance. An empirical Bayes framework is then used to improve the accuracy of these estimates using a linear regression model; the observed means and variances from each site are modelled as a combination of the harmonised mean and variance along with additive and multiplicative effects. When fitting this model, the harmonised mean and variance values can be obtained for a given feature, which allows the whole distribution of feature values to be realigned, thus removing site-effects in the process. Factors such as age and sex can be included as covariates in the model to preserve important biological trends, if subjects scanned in multiple sites are not matched. In (Fortin et al. 2017), when compared to RAVEL (see section 2.2.2.2), it was found that ComBat reduced more the variability in measures for several diffusion derived measures. Unlike some of the previously mentioned regression models where the correction needs to be performed for individual features separately, ComBat can work on finalised parameter maps which allows for the adjustment of scanner and site differences across all features collectively.

Figure 2.4 gives a visual illustration of of how ComBat takes unharmonised distributions of feature values and realigns them. The observed multi-site

feature values are y_{ij} measured in volume of interest (VOI) j and scanner i, α is the average value of the feature, γ_i is an additive scanner effect, δ_i is a multiplicative site effect ϵ_{ij} is the error term.

the harmonised feature values are given by

$$e_{ijf} = \frac{y_{ij} - \hat{\alpha} - \hat{\gamma}_i}{\hat{\delta}_i} + \hat{\alpha} \tag{2.3}$$

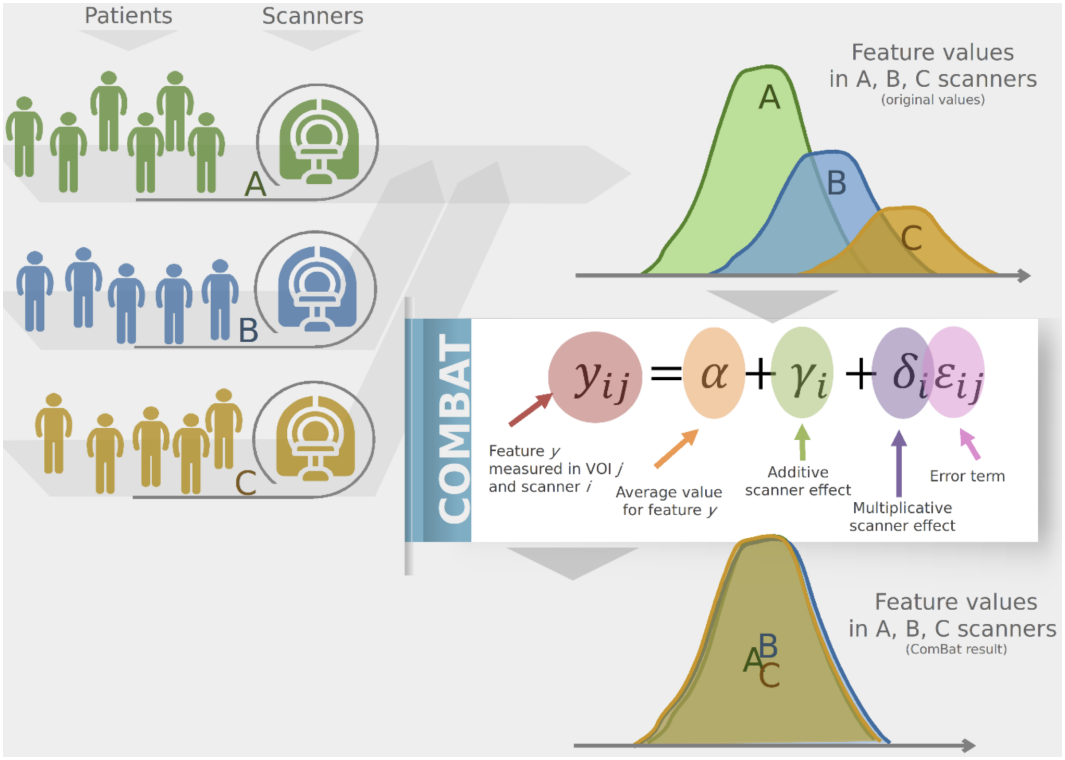

Figure 2.4: *Harmonisation using the ComBat method that realigns the distributions of feature values. Figure extracted from (Orlhac et al. 2021).*

Since the inception of ComBat, several approaches have been developed which expand on the original work. It has been noted that while accounting for site removes unwanted variation in the data, there is the possibility that it may remove relevant information. In an effort to address this, the authors in (Wachinger et al. 2021), augmented the original ComBat model to explicitly preserve or remove variables of interest. It is noted that besides known effects such as scanner manufacturer or magnetic field strength, *unknown site effects* can also be present, which cannot be explicitly represented, but have

an effect on differences across sites/scanners. To account for such unobserved effects, principal components are computed across all imaging features. This approach is common in genome-wide association studies (GWAS) where principal components are added to the regression model to remove unobserved subpopulation structures within the sample. To assess whether this approach was successful at removing dataset-specific information as well as relevant biological information, the authors used the fact that ventricles grow with age while atrophy of the hippocampus increases. If biologically relevant information was preserved, then this correlation would be observed in the harmonised data. Results showed that this developed version of ComBat achieved less than 50% classification accuracy in *name that dataset game* (a random forest classifier was used) while indeed maintaining the positive correlation of lateral ventricle volume with with age and the negative correlation of hippocampus volume with age. This correlation was present in the unharmonised data and therefore confirmed to a reasonable degree the preservation of biologically meaningful measures of interest.

In (Chen et al. 2020), the authors note that the ComBat model does not address potential covariance in scanner effects. They therefore proposed a method called Correcting Covariance Batch effects (CovBat) for removing scanner effects in mean, variance of each site and also covariance between sites. The concept involves shifting covariance matrices of individual scanners (covariance of each scanner was set as the sample correlation matrix of cortical thickness observations) to the general covariance across sites in a similar way to how normal ComBat modifies observations to bring the mean and variance of each scanner to the pooled variance across sites. In that particular study, the covariances were calculated for cortical thickness measures. Applying CovBat significantly decreased the differences between covariance matrices across scanners. The results showed a scanner classification accuracy of close to chance in a Monte-Carlo split-sample experiment for prediction of scanner

manufacturer labels. It was additionally shown that CovBat preserved biological associations by using a random forest to predict whether the harmonised data could correctly detect Alzeimer's disease status and sex.

Another development of the original ComBat is ComBat-GAM (Pomponio et al. 2020). ComBat-GAM aims to capture nonlinearities in age-related differences in brain structure. A generalised additive model (GAM) is integrated within the originally proposed ComBat. GAMs allows for the addition of predictors or features to enter a model in an additive way. The key difference to the regular form of ComBat is that predictors can be wrapped in a function with a significant amount of complexity and nonlinearity. The results showed ComBat-GAM to be effective at removing bias and variance in volumes of selected ROI's, total grey matter volume and total white matter volume. It was additionally shown that ROI volumes harmonised by ComBat-GAM outperformed ComBat using a linear model in an age prediction task. The improved performance of GAMs over linear models confirms the existence of nonlinear trends in brain structures.

A limitation of ComBat and its derivatives is pointed out in (Karayumak, Bouix, Ning, James, Crow, Shenton, Kubicki & Rathi 2019) where the authors state that the assumption that site-effect parameters follow a particular parametric prior distribution may not generalise to all scenarios or measures. This limitation is a specific example of the limitations of regression-based approaches in general. Scanner effects and how these affect features have to be defined a-priori to parameterise effects in the regression models. This may be a good first order approximation but it is unclear how generalisable it is. The authors also note that it is unclear how nonlinearities in the signal caused by site-effect propagate through the model fitting procedures.

2.2.3.3 Regression-based approaches (ROI-based)

Regression-based approaches can also be applied on the ROI level. An example is in (Pagani et al. 2010) who conducted a study to assess whether a group of multiple sclerosis patients could be distinguished from a group of healthy controls with data originating from 8 different sites having different scanners. Their analysis was based on DTI-derived metrics (FA, MD, Dax & Drad) in regions of interest (ROIs) within the Corpus Callosum. Their approach was to use an analysis of variance (ANOVA) test to determine the independent influence of factors, such as scanner manufacturer, on the results of derived features and then to correct for those factors. The aims of the authors were simply to be able to correct data so that patients and healthy subjects could be distinguished, and this method proved to be successful in certain regions. This is a good starting point, but the wider aim of harmonisation is to remove site related variability while preserving biological variability. The regression method used in this study did not fulfil that aim entirely since it only went as far as to separate healthy controls from subjects and it managed this only in specific ROIs.

In (Venkatraman et al. 2015), regression based approaches are improved upon. In a study involving 4 different scanners with different field strengths, they were able to reduce the variability in a large number of ROIs for diffusion measures (FA and MD). The difference in this method is that it was data-driven. A linear mixed effects model trained on data from 544 participants was used to learn scanner differences in different regions. The regression model was used to estimate differences in measurements due to scanner effects and this was used to calculate a correction factor, which was then applied to each ROI to remove the effects. A limitation of applying a single correction factor to each ROI is that the site effects within an ROI are assumed to be constant. This may be a harmless assumption to make for structural modalities in sta-

ble areas but, as (Mirzaalian et al. 2016) point out, this may be inadequate for analysis of tractography results where tracts travel between distant regions.

ComBat has also been applied on functional connectivity values obtained from resting-state fMRI data. In (Yu et al. 2018), the connectivity values between two ROIs were given as inputs to the model which was used to remove additive and multiplicative site effects. Before performing ComBat harmonisation on the functional connectivity matrices, all the network connectivity values (estimated by Pearson correlation) displayed statistically significant site effects. After the harmonisation, there were no remaining statistically significant site effects. (Yamashita et al. 2019) sate that a potential limitation of ComBat is that site differences are estimated without taking into account sampling bias (caused by sampling from among different subpopulations) which could result in biologically meaningful sampling bias being removed.

To address this, (Yamashita et al. 2019) suggested a travelling-subjects approach for the harmonisation of resting state fMRI data. This allowed a linear regression model to be used which estimated site-related bias separately from sampling bias. The site-related bias was quantified as the deviation of the connectivity value for each functional connection from its average across all sites. This average was determined using travelling subject data and to harmonise the data, the measurement bias was subtracted from the connectivity values. Their results indicate that this method removed measurement bias and also improved signal-to-noise ratios.

2.2.3.4 Machine/Deep learning

Methods like ComBat, as noted by (Garcia-Dias et al. 2020), require sample sizes which are large enough to be statistically representative of each scanner included in the study. These conditions could prove a hindrance to clinical application where assessments and predictions are likely to be made from sin-

gle images and from scanners that are not guaranteed to have been part of an initial training set. To address this, the authors propose Neuroharmony (Garcia-Dias et al. 2020), a tool for harmonising unseen data which they apply on data from T1-weighted images. 15,026 subjects were used to train a machine learning tool to learn the relationship between image quality metrics (IQMs) and ComBat-harmonised brain volumes. The image quality metrics, such as contrast between white matter and grey matter and signal to noise ratio, are shown to be associated with the scanner used to acquire images. The mapping between IQMs and corrected volumes, learned using a random forest, are used to predict features of an image using as inputs IQMs from an unseen scanner. This is widely generalisable as the chosen IQMs are directly measurable from individual MRI images. The tool was successful in removing scanner related bias in brain volume measurements in 101 ROIs. One of the limitations of the method is that the accuracy of the method differs between regions. Specifically, regions showing greater variability prove to be more difficult to harmonise. An additional limitation is that if the value of particular IQMs falls out of the range of the training sample used, effective harmonisation cannot be guaranteed.

A more recently developed deep learning approach is presented in (Dinsdale et al. 2021). The authors leverage a deep learning technique known as domain adaptation. The aim of domain adaptation is to find a representation of features invariant to domain which is subsequently linked to a task of interest. If the domain is made to be an MRI scanner, this can be adapted to a harmonisation problem and the task can be a specific feature extraction process performed by a feature extractor θ_{repr} (See Figure 2.5). In the example presented the task was structural image segmentation into predefined classes (WM, GM and CSF). This is performed by a label predictor with parameters θ_p. The next component of the architecture is a domain classifier with parameters θ_d, which aims to predict were the data came from. The network is trained

by minimising the loss on the main task, which in this case is segmentation label prediction, while maximising the loss on the domain/scanner classification. The harmonisation is completed when it is no longer possible to predict which scanner the data came from, while at the same time achieving maximum segmentation accuracy. To assess the quality of the harmonisation the Dice score was computed between the generated segmentations and segmentations performed by FMRIB's Automated Segmentation Tool (Zhang et al. 2001) which served as a proxy for manual segmentation. A Dice score of 0.91 was achieved and with a scanner classification accuracy of 51% indicating highly accurate segmentation with results agnostic to the scanner on which they were acquired.

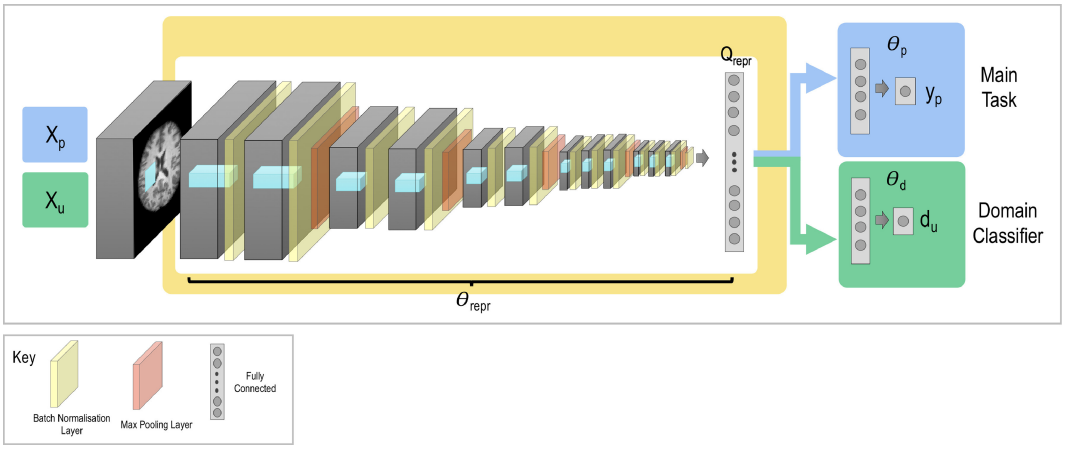

Figure 2.5: *General network architecture. The network is formed of three sections: the feature extractor with parameters θ_{repr} the label predictor with parameters θ_p and the domain classifier with parameters θ_d. X_p represents the input data used to train the main task with labels y_p and X_u represents the input data used to train the steps involved in unlearning scanner with labels d. Figure extracted from (Dinsdale et al. 2021).*

2.2.4 Summary of Existing Methods

We saw a range of methods that have been proposed to solve the harmonisation challenges from different angles. We summarise the reviewed methods in Figure 2.6, indicating which imaging modalities they have mostly been applied on. Even if this is not an exhaustive list, it provides a representative picture. We can see that post-acquisition harmonisation has been explored mostly for

structural and diffusion modalities and efforts on SWI and functional data are much less explored. Nevertheless, it is also clear that harmonisation is still an open challenge and no comprehensive solution exists at the moment. Generalisability to new features/modalities is far from trivial, while evaluation of the various approaches seem to be based on ad-hoc, study-specific criteria, which may not always be objective. We discuss relevant challenges in the following sections and how the current thesis aims to contribute towards these directions.

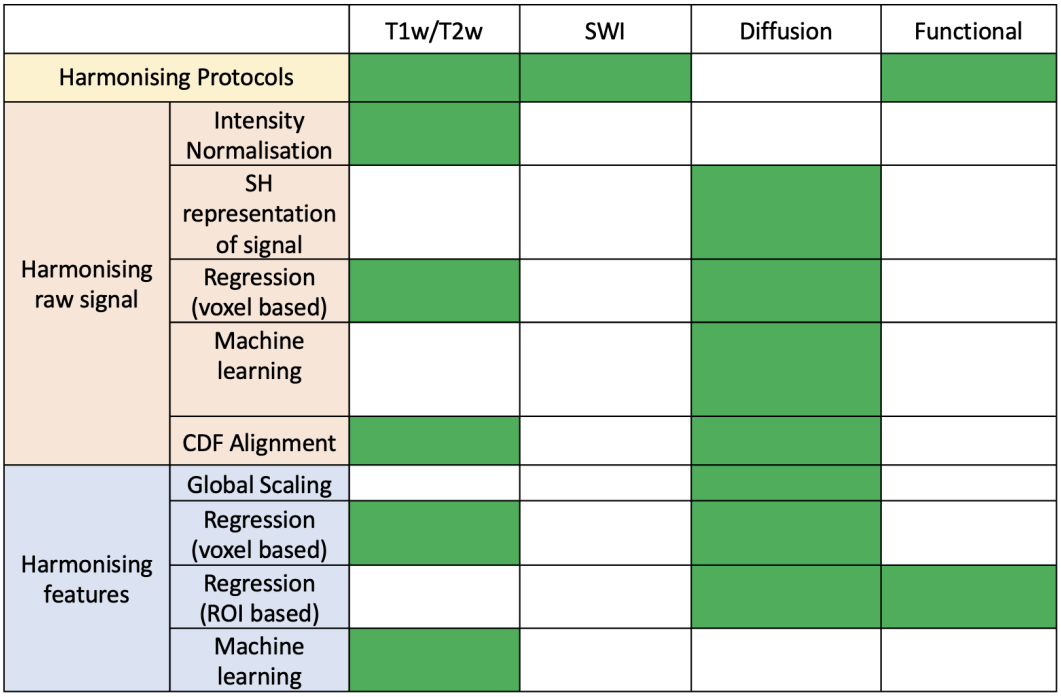

Figure 2.6: *Summary table showing different modalities and which groups of reviewed methods have worked on harmonising them.*

2.3 Thesis Aims

2.3.1 Overcoming the Challenge of Evaluating Harmonisation Approaches

Harmonisation aims to remove unwanted variability induced by scanner or site effects while preserving true biological variability. The previous section highlighted a number of approaches which endeavour to harmonise data. There is

however a challenge in evaluating these approaches: there is a lack of a consistent reference and criteria for assessing the harmonisation results against.

One of the simplest methods that has been used is to use the bias and variance in the distribution of imaging derived metrics derived from travelling subjects, or human phantoms, scanned in different scanners. A scanner-specific correction factor is inferred from the human-phantom data and then applied to feature maps from other scans. This was the method used in (Pohl et al. 2016) where it was seen that the distributions of diffusion derived metrics after harmonisation showed a higher degree of overlap than before harmonisation. This approach serves as good starting point for ensuring groups of data points come from the same distribution however, it is limited because it fails to assess the extent of harmonisation on the individual feature level. A further limitation is that, due to an absence of a ground truth, the harmonisation of data points is assessed with respect to each scanner. For some applications this may be enough but other applications may require a more concrete reference point especially since scanners differ in their ability to produce consistent results (Vollmar et al. 2010).

In (Vollmar et al. 2010), 2 within-scanner repeats are used as a reference for assessing the methods reliability. The smaller the difference between the coefficient of variation (or the larger the correlation) of between-scanner measurements and within-scanner measurements the more effective the harmonisation approach is deemed to be. Although this approach has a "gold standard", there are only 2 within-scan repeats which is relatively low. A similar approach with more within-scan repeats per subject would be ideal, so that within-scanner variability can be assessed and used as a reference/target for post-harmonised between-scanner variability.

In addition, a number of studies used the approach of subject matching for

evaluations in the asbence of a gold standard. For example, in (Mirzaalian et al. 2016), where raw diffusion signal is harmonised, it is assumed that two separate groups of individuals matched for age, gender, handedness and socio-economic status should have similar diffusion profiles and any difference is attributed to scanner-related inconsistencies. A similar method is is used in (Fortin et al. 2017) to harmonise derived FA and MD maps from diffusion data were subjects are matched for age, gender, ethnicity, and handedness. A limitation of this approach is that a possibility remains that observed differences between sites may simply reflect differences in participant characteristics.

In this work, we aim to solve the evaluation challenge in two ways: a) Provide within-scanner variability references for repeated measurements. We scan a number of subjects 6 times in the same scanner and we use these within-scanner variability metrics as lower bound for between-scanner variability. b) To avoid challenges with subject matching we use a travelling heads pradigm. The same subjects scanned in 6 different scanners to provide metrics of between-scanner variability for the same anatomical, functional, microstructural features.

2.3.2 Building a Comprehensive Travelling Heads Dataset

As pointed out in (Badhwar et al. 2020) "only a single cohort experiment can unambiguously capture inter-site differences, with the same individual(s) being scanned repeatedly at each site". A number of previous studies have followed this paradigm, summarised in Table 2.1.

Specifically, in (Hawco et al. 2018) 4 participants were scanned in 5 scanners (2 Siemens Prisma, 1 Siemens Tim Trio, 1 GE 750w Discovery & 1 GE 750 Signa). T1, dMRI and rfMRI data were collected, but no within-scanner repeats were acquired. In (Tax et al. 2019), 14 participants were scanned in 3 scanners (1 Siemens Prisma, 1 Siemens Connectom and 1 GE Signa Ignite).

Table 2.1: Summary of travelling heads data set.

Dataset/Study	No. of scanners	Scanners Vendors	Modality	No. of subjects: between scanner	No. of subjects: within scanner
(Hawco et al. 2018)	5	Siemens(3) GE(2)	T1 dMRI rfMRI	4	N/A
(Tax et al. 2019)	3	Siemens(2) GE(1)	T1 dMRI	14	N/A
(Kurokawa et al. 2021)	4	Siemens(4)	T1 dMRI	9	4
(Duff et al. 2021)	4	Siemens(3) GE(1)	T1, T2 SWI dMRI rfMRI ASL	8	N/A
(Tanaka et al. 2021)	12	Siemens(7) Philips(3) GE(2)	T1 rfMRI	9	N/A
Our Dataset	6	Siemens(3) Philips(2) GE(1)	T1, T2 SWI dMRI rfMRI	10	4

T1 and dMRI data were collected and there are no within-scanner repeats acquired. In (Kurokawa et al. 2021), 9 participants were scanned in 4 scanners (1 Siemens Prisma, 1 Siemens Prisma fit, 1 Siemens Skyra fit, 1 Siemens Verio). T1 and dMRI data were collected and within-scanner repeats were acquired for a separate 4 subjects. In (Duff et al. 2021) 8 participants were scanned in 4 scanners (3 Siemens Prisma Scanners, 1 GE scanner). T1, T2 FLAIR, dMRI, SWI, ASL and rfMRI data were acquired and no within-scanner repeats were acquired. In (Tanaka et al. 2021) 9 participants were scanned in 12 scanners (1 Siemens Trio, 1 Siemens TimTrio, 1 Siemens Skyra, 2 Siemens Verio, 1 Siemens VerioDot, 1 Siemens Spectra, 3 Phillips Achieva, 1 GE Signa HDxt & 1 GE MR750W). T1 and rfMRI data have been acquired, but no within-scanner repeats were acquired.

Among these studies we see that they typically cover one or two of the major MRI vendors (Kurokawa et al. 2021, Duff et al. 2021), with 2-4 scanners

used in total (Tax et al. 2019). Furthermore the acquired images reflect either one or maximum of two MRI modalities (Tax et al. 2019). For instance, the dataset presented in (Tax et al. 2019), is limited to dMRI and T1 modalities and only 3 scanners from 2 different vendors are used. Similarly, the SRPBS Travelling Subject MRI dataset (Tanaka et al. 2021) consists of imaging data acquired from all three major vendors but is limited to rfMRI and T1 modalities. Although the dataset acquired in (Hawco et al. 2018) spans a broader range of modalities by acquiring structural, diffusion and functional data, it is limited by the relatively low number of subjects used (N=4) as well as restricted coverage of MRI vendors.

We therefore aimed to acquire a dataset, which is more comprehensive than before in a number of ways. It includes a) scanners from all vendors and from two generations within each vendor, b) within-scanner repeats to allow for relevant evaluations, c) five different imaging modalities in 10 subjects, d) using scanning protocols that are aligned with one of the few population-level resources, the UK Biobank.

2.3.3 Mapping the Need for Harmonisation for Thousands of Multi-modal Imaging-derived Features

Existing approaches have explored the need for harmonisation over a very limited set of features, including mostly cortical thickness and volumes and DTI metrics, such as FA and MD. However, there are many features that can be routinely extracted from multi-modal images.

Furthermore, open questions exist, for instance which imaging modalities are more/less prone to between-scanner effects? And which features extracted from these modalities are more prone to these site effects? In addition, for many features there are different processing steps, models and pipelines to

obtain them from the same data. For example, when extracting the volumes of subcortical regions one has the option of FreeSurfer-based segmentation or registration-based segmentation. Another example is found when extracting DTI-based metrics. One can extract these metrics in ROI's defined using an atlas or using tractography. Given this, which is the way to derive these features that it is more immune to site effects? In this thesis, we will explore such questions for thousands of multi-modal imaging-derived features.

2.3.4 Testing Harmonisation Approaches

While a number of harmonisation approaches have been developed, what's missing is objective ways to evaluate them. This thesis addresses this by identifying the optimal pipelines and processing steps that minimise between-scanner variability in extracted features and also explicitly testing the performance of post-processing harmonisation tools and checking whether the harmonised features between-scanners are indeed less variable.

This analysis is enabled by the travelling heads data set which we have acquired and what we demonstrate is merely the beginning of the variety of investigations that can be carried out to assess the multitude of harmonisation approaches which currently exists and are yet to be developed.

Chapter 3

A Multi-modal Travelling-heads Harmonisation Resource for Brain MRI

In the previous chapter, a number of methods and studies that attempt to provide harmonisation solutions for neuroimaging data were presented. Most of the previous approaches are unimodal, they are limited in the range of scanners they consider and they typically lack explicit gold standards to compare the harmonisation results against. In this chapter we present the building of a resource that aims to address the above challenges. Following a travelling-heads paradigm (healthy volunteers scanned repeatedly across multiple sites) we acquired data from 6 scanners, 5 sites and 5 neuroimaging modalities. The scanners include systems from all vendors that span different within-vendor generations. To obtain ground-truth on scan-rescan variability, within-scanner repeats of the same subjects have also been acquired. We present in this chapter the main setup with acquisition protocols, along with multi-modal quality control procedures. The data acquired will be publicly released in a repository providing a resource for the community working in development of harmonisation methodology.

3.1 Personal contributions

My personal contribution to the work performed in this chapter came after the protocols parameters had already been adapted to match those of the UK Biobank as closely as possible. I contributed to the subsequent data acquisition and to modifying protocols to resolve issues that were unknown in earlier testing (Section 3.7)

3.2 Introduction

As reviewed in the previous chapter, there have been multiple recent attempts for harmonising multi-site neuroimaging data. However, studies have focused on single modalities at a time, while they typically lack objective ways and datasets to compare the post-harmonisation results and evaluate them.

Several prominent initiatives such as the UK Biobank (Miller et al. 2016), ADNI (Jack Jr et al. 2008), ABIDE (Di Martino et al. 2014) and ABCD (Casey et al. 2018) have provided large human brain MRI datasets and have acted as testbeds for the evaluation of harmonisation methods (Beer et al. 2020, Dinsdale et al. 2021). Although these datasets comprise of a large number of participants, they are limited in a number of key ways in being used for the assessment of harmonisation methods. One of these limitations is that, even if they comprise of multi-site data, different participants were recruited at each site. This keeps open the possibility that site effects simply reflect differences in participant characteristics. On the other hand, having a single cohort study can unambiguously capture inter-site differences, with the same individual(s) being scanned repeatedly at each site (Badhwar et al. 2020).

It has therefore been demonstrated that a travelling-heads (or “human phan-

tom") paradigm is well suited to evaluating harmonisation approaches (Maikusa et al. 2021, Yamashita et al. 2019). These are datasets wherein multiple participants travel and get scanned at multiple imaging sites. By controlling for participant effects between sites, the effects of scanner induced bias and variance can be assessed.

A number of travelling-heads datasets have been previously collected (Tax et al. 2019, Kurokawa et al. 2021, Hawco et al. 2018, Duff et al. 2021). Typically however, they have been limited in covering only one or two of the major MRI vendors (Kurokawa et al. 2021, Duff et al. 2021) , or with 2-3 scanners used in total (Tax et al. 2019) . Furthermore the acquired images reflect either one or a maximum of two MRI modalities (Tax et al. 2019). For instance, the dataset presented in (Tax et al. 2019), is limited to dMRI and T1 modalities and only 3 scanners from 2 different vendors are used. The SRPBS Travelling Subject MRI dataset (Kurokawa et al. 2021) consists of imaging data acquired from all three major vendors but is limited to rfMRI and T1 modalities. Although the dataset acquired in (Hawco et al. 2018) spans a broader range of modalities by acquiring structural, diffusion and functional data, it is limited by the relatively low number of subjects used (N=4) as well as restricted coverage of MRI vendors.

In light of these limitations of existing studies, the resource presented here aims to be more comprehensive in the following ways: i) by acquiring data from all 3 major vendors and from different generations of scanners from the same vendor, ii) by acquiring data from physically different imaging sites, where radiographers and practices are different, reflecting a closer-to-reality-scenario, iii) by acquiring data from many neuroimaging modalities using modern acquisition protocols and capturing standard anatomical MRI (T1 and T2-weighted), microstructural and connectivity info (diffusion and susceptibility-weighted MRI) and functional networks (resting-state functional MRI), iv) by acquiring data

that allows the assessment of within-scanner, within-subject scan-rescan variability in addition to between-scanner variability.

This chapter gives an overview of the setup we used, including scanners, sites, subjects for which different data were acquired. An overview of the imaging modalities is also given including protocol details and sequence information. Quality control (QC) comparisons across the scanners and modalities are also presented with explorations on how quality differs between scanners and vendors due to specific hardware features. The data presented here will be anonymised and publicly released to the community, providing a resource for future developments on neuroimaging harmonisation.

3.3 Methods

3.3.1 Acquisition Strategy

Multi-modal neuroimaging data were obtained from $N = 10$ travelling heads, i.e. healthy subjects that were scanned across multiple scanners and sites. We also acquired an additional 5 repeat scans of some subjects ($M = 4$) on the same scanner, to assess scan-rescan within-scanner variability. Figure 3.1 gives an overview of the overall acquisition strategy.

The data were acquired using six 3.0T scanners physically located at 5 different imaging sites in Nottingham and Oxford. These scanners span all the 3 major MRI vendors (Siemens, Philips and GE) and also capture some within-vendor variability across older and modern systems (for instance Philips Achieva vs Ingenia, Siemens Trio vs Prisma). The technical specifications of each scanner are given in Tables 3.1 and 3.2 . As can be seen there is intentional variability in the type of coil channels used (32 vs 64), in the maximum gradient strength (40 to 80 mT/m) and bore size (narrow vs wide bore), capturing differences

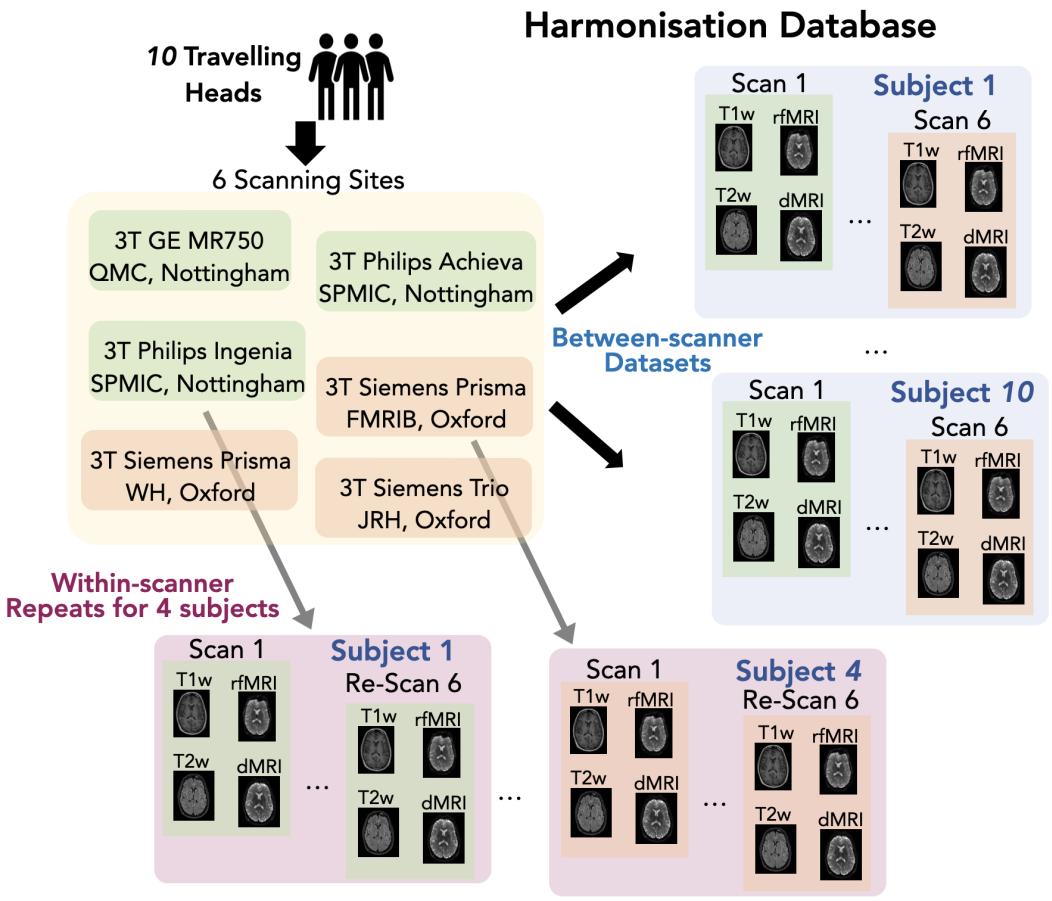

Figure 3.1: *Overview of overall acquisition strategy. Multi-modal neuroimaging data were obtained from healthy participants scanned in 6 different scanners. Each scanning session comprised of multiple imaging modalities, including T1 and T2-weighted, diffusion and resting-state functional MRI and susceptibility-weighted imaging. The scanners were all 3.0 Tesla and physically located in 5 imaging sites: i) The SPMIC-QMC in Nottingham, ii) The SPMIC-UP in Nottingham, iii) The WIN-FMRIB at the John Radcliffe Hospital in Oxford, iv) The WIN-OHBA at the Warneford Hospital (WH) in Oxford and v) The OCMR at the John Radcliffe Hospital (JRH) in Oxford.*

that would be anticipated in a real-world scenario.

3.3.2 Participants and Ethical Approvals

The data were obtained from $N = 10$ healthy participants (mean age 34 $\pm$ 9.4 years; 8 male, 2 female. Each subject was scanned once at each of the 6 scanners. For a 4 of the subjects, an additional 5 repeat scans of the subject in the same scanner were acquired (see Table 3.3). It was initially attempted to acquire data from each subject in the shortest possible time frame. Due to the pandemic, lockdown restrictions and relevant disruptions, data collection

Table 3.1: The location and technical specifications for each scanner.

Scanner	Field Strength	Bore Size [cm]	Max Gradient Strength [mT/m]	Gradient Slew Rate [T/m/s]	Coil number of Channels
3T Siemens Prisma FMRIB	3T	60	80	200	32
3T Siemens Prisma WH	3T	60	80	200	64
3T Siemens Trio JRH	3T	60	45	200	32
3T Philips Ingenia SPMIC	3T	70	45	200	32
3T Philips Achieva SPMIC	3T	60	Dual: 40 (80)	200 (100)	32
3T GE MR750 QMC	3T	60	50	200	32

Table 3.2: The software Version for each scanner.

Scanner	Software Version
3T Siemens Prisma FMRIB	syngo MR E11
3T Siemens Prisma WH	syngo MR E11
3T Siemens Trio JRH	syngo MR B17
3T Philips Ingenia SPMIC	5.3.1/5.3.1.0 (subject 03268, 03997, 13192, 13305, 14229, 14230) and 5.6.1/5.6.1.1 (subject 10975, 12813, 14221, 14482)
3T Philips Achieva SPMIC	5.3.0/5.3.0.3 (subject 03268, 03997, 13192, 13305, 14229, 14230, 14482) and 5.6.1/5.6.1.1 (subject 10975, 12813, 14221)
3T GE MR750 QMC	DV24.0 R02 1607.b

ended up being particularly challenging and lasted more than intended. Across the different scanners, scans for each subject were completed within a period of between 203 and 671 days. The median time to acquire all the scans across different sites was 438 days. To the contrary, the median time to acquire all within-scanner rescans was 88 days. Details for each participant are presented in Table 3.3.

The scans were performed under two ethics protocols for healthy volunteers at Nottingham (PI: Sotiropoulos, Ethics: FMHS-36-1220-03 H14082014/47) and Oxford (PI: Jenkinson, Development Ethics). Informed consent was obtained

Table 3.3: Participant information including time taken to acquire all scans.

Subject ID	Sex	Age	Within-scan repeats	Between-scan Interval (days)	Within-scan Interval (days)
03286	M	48	No	492	N/A
03997	M	37	No	266	N/A
10975	M	25	No	455	N/A
12813	F	24	No	562	N/A
13192	M	47	Yes	314	38 (Prisma FMRIB)
13305	M	42	No	671	N/A
14221	M	25	No	555	N/A
14229	M	35	Yes	298	83 (Prisma WH)
14230	F	25	Yes	203	92 (Trio)
14482	M	24	Yes	500	388 (Achieva)

from all volunteers. Scan time costs were provided in part by the Nottingham Biomedical Research Centre, by the SPMIC-School of Medicine PhD student and scan time allocation fund and by the WIN Centre. Scanners were operated by local radiographers and physicists (Mr Jon Campbell for FMRIB Prisma, OCMR Trio and FMRIB OHBA, Mr Andrew Cooper for SPMIC-QMC, Dr Olivier Mougin and Prof Paul Morgan for SPMIC-UP Philips Ingenia and SPMIC-UP Philips Achieva).

3.3.3 Overview of Multi-modal Acquisition Protocols

As a reference, we followed a modern, yet not bespoke (and therefore applicable to older scanners), multi-modal neuroimaging protocol, as provided by the UK Biobank Imaging study ((Miller et al. 2016). This represents a good compromise between richness in features extracted (five imaging modalities), image quality robustness and acquisition time (relatively short scan time of 30-40 minutes). Also, a population-level cohort (40,000 subjects) is currently available with this protocol that allows assessments of biological variability. The acquisition protocols for each scanner followed closely the original UK Biobank protocols (developed for a Siemens Skyra), respecting however the good practice policies for each scanner that can render between-vendor differences to achieve optimal quality in each scanner. For instance, matching

blindly acceleration factors for SENSE, GRAPPA and ARC may not always be optimal for data quality as these parallel imaging reconstructions and their implementations across vendors can have different artifacts and behaviour.

We used 5 modalities out of the UK Biobank neuroimaging protocol. These include: i) **T1-weighted** imaging, a technique which exploits the differences in T1 relaxation times of tissues to probe overall morphology and show strong contrast between grey and white matter. ii) **T2-weighted FLAIR** structural images, rely on the differences in tissue T2 relaxation times. The contrast between grey and white matter is subtle relative to T1. Its strength lies in its utility in depicting pathologies such as white matter lesions. iii) **Susceptibility-weighted** imaging (SWI), a technique which purposely enhances the effect of local field variations caused by magnetised tissue constituents, such as iron content or calcium content. SWI enhances the appearance of veins due to their inherent deoxyhemoglobin content. iv) **Diffusion-weighted** imaging, sensitive to the anisotropic nature of thermally-driven motion of water molecules in biological tissue and can be used to probe tissue micro-structure. Diffusion metrics (such as DTI) can inform of the integrity of the tissue microstructure and derived orientation estimates in white matter can be utilised in tractography algorithms to give information about the connectivity of brain regions. v) **Resting-state functional** MRI, which is sensitive to changes in blood oxygenation linked to neuronal activation, can be used to study intrinsic oscillatory activity in the brain (i.e. at "rest", in the absence of stimuli) and extract functional networks. Synchronous fluctuations in different regions of the brain are indications of regional activations and therefore indirect evidence of regions communicating with each other as part of the same network.

In addition to these modalities, we also acquired blip-reversed spin-echo for the generation of fieldmaps used to carry out corrections for susceptibility-induced artefacts in the EPI acquisitions (Andersson et al. 2003).

3.3.4 Acquisition Modifications to Accommodate all Scanners

The original acquisition protocol was developed by the UK Biobank consortium for Siemens Skyra scanners. Where possible, we were able to exactly match parameters to those in the UK Biobank (for instance spatial resolution, field of view, contrast timings, temporal and angular resolution in functional and diffusion MRI, number of echoes and echo times for SWI) but there were also times when this was not possible due to inherent scanner/vendor differences (e.g. maximum gradient strength or lack of simultaneous-multislice reconstructions). In these situations, we strived to reach a reasonable compromise. In this section we describe the acquisition protocol for each modality, including explanation in some of the instances where we deviated from the original protocol for certain scanners. For the majority of these options the choices were a balance between: a) How close we could get to the UK Biobank, b) How far we could go from the standard practice on each scanner. Pushing the former over the latter would in theory create more "harmonised" protocols to start with, but in practice this can either induce suboptimal quality in some scanners or create an artificial level of consistency not anticipated in real-world scenarios when pulling multi-site data together. We therefore attempted to fulfil (a), without however compromising (b).

T1-weighted

Each subject underwent a magnetisation-prepared rapid gradient-recalled echo (MPRAGE) T1-weighted scan. The protocol details are presented in Table 3.4. A high-resolution (1 mm^3), whole-brain scan was obtained for all scanners. Gradient distortion correction (GDC) was turned off for the Siemens scanners because the Siemens on-scanner corrections have been found to give

inconsistent results, particularly for 2D EPI acquisitions (scanner-corrected 3D and 2D acquisitions of the same subject cannot be successfully aligned with a rigid body transformation). As a result, these corrections were applied in the post-imaging processing pipeline for Siemens scanners using a proprietary file that characterises the spatial distribution of gradient nonlinearities. To our knowledge, GDC functioned correctly for the non-Siemens scanners so this was performed on the scanner. This applies for all other modalities we acquired. For the Philips and GE scanners we used the provided on-scanner GDC correction option, following best practice in the respective sites. Vendor-provided pre-scan normalise was used for all scanners. Scan time was in the order of 5 minutes for this modality.

Table 3.4: Protocol details for T1-weighted acquisition across the different scanners. The reference UK Biobank protocol is shown on the top row.

Scanner	Sequence	Voxel Size	Matrix Size	TR/TE	Coil of Channels	Parallel imaging Factor	Partial Fourier	Pre-scan Normalise	GDC
3T Siemens Skyra (UK Biobank)	3D - MPRAGE - sagittal	1mm^3	208x256x256	2000/2.01	32	2	Off	Yes	Off
3T Siemens Prisma FMRIB	3D - MPRAGE - sagittal	1mm^3	208x256x256	2000/2.03	32	2	Off	Yes	Off
3T Siemens Prisma WH	3D - MPRAGE - sagittal	1mm^3	208x256x256	2000/2.03	64	2	Off	Yes	Off
3T Siemens Trio JRH	3D - MPRAGE - sagittal	1mm^3	208x256x256	2000/1.97	32	2	Off	Yes	Off
3T Philips Ingenia SPMIC	3D - MPRAGE - sagittal	1mm^3	208x256x256	7000/3.2	32	2	Off	Yes	On
3T Philips Achieva SPMIC	3D - MPRAGE - sagittal	1mm^3	208x256x256	7100/3.2	32	2	Off	Yes	On
3T GE MR750 QMC	3D - BRAVO - sagittal	1mm^3	176x256x256	6600/2.9	32	2	Off	Yes	On

T2-weighted

With the exception of the GE MR750, all the T2-weighted scans were performed using a 3D T2-weighted-Fluid-Attenuated Inversion Recovery (T2w FLAIR) sequence that allowed high-resolution data (almost 1 mm^3 isotropic). The protocol details are presented in Table 3.5 The MR750 did not have 3D T2-weighted FLAIR functionality (could either provide a 3D FLAIR with no T2-weighting or a 2D T2w FLAIR). We instead ran a 2D T2-weighted FLAIR, but we had to compromise with spatial resolution due to timing constraints. Scan time was in the order of 8 minutes for the GE MR750 and 4 minutes for the rest of the scanners. Same GDC and pre-scan normalise options were followed as before.

Susceptibility-weighted imaging

The SWIs were acquired using anisotropic, complex data for 2 echoes, roughly matching around $TE_1 \sim 9s$ and $TE_2 \sim 20s$. The GE scanner software (SWAN sequence) acquired 7 echoes and the two echoes closer to TE_1 and TE_2 were extracted during processing. For the Siemens scanners, individual coils were saved separately to enable combination of phase images, and they were combined in post processing whereas for the non-Siemens scanners, these were combined on the scanner. Magnitude and phase images were saved for all the scanners. Scan times were in the order of 2.5 minutes for all scanners.

Table 3.5: Protocol details for T2-weighted FLAIR acquisition across the different scanners. The reference UK Biobank protocol is shown on the top row.

Scanner	Sequence	Voxel Size	Matrix Size	TR/TE	TI	Parallel imaging Factor	Partial Fourier	Pre-scan Normalise	GDC
3T Siemens Skyra (UK Biobank)	3D - T2w FLAIR - sagittal	1.05mm× 1mm × 1mm	256x288x256	5000/395	1.8	2	87.5%	Yes	Off
3T Siemens Prisma FMRIB	3D - T2w FLAIR - sagittal	1.05mm× 1mm × 1mm	192x256x256	5000/397	1.8	2	87.5%	Yes	Off
3T Siemens Prisma WH	3D - T2w FLAIR - sagittal	1.05mm× 1mm × 1mm	192x256x256	5000/397	1.8	2	87.5%	Yes	Off
3T Siemens Trio JRH	3D - T2w FLAIR - sagittal	0.99mm× 1mm × 1.05mm	192x256x256	5000/396	1.8	2	87.5%	Yes	Off
3T Philips Ingeia SPMIC	3D - T2w FLAIR - sagittal	1.05mm× 1mm × 1mm	192x256x256	5000/322	1.65	2 (AP), 2 (RL)	Off	Yes	On
3T Philips Achieva SPMIC	3D - T2w FLAIR - sagittal	1.05mm× 1mm × 1mm	192x256x256	5000/290	1.65	2 (AP), 2 (RL)	Off	Yes	On
3T GE MR750 QMC	2D - T2w FLAIR - sagittal	1.05mm× 1.05mm× 1.05mm	160x160x116	12000/95	2.71	2	Off	Yes	On

I

Table 3.6: Protocol details for SWI acquisition across the different scanners. The reference UK Biobank protocol is shown on the top row.

Scanner	Sequence	Voxel Size	Matrix Size	TE [msec]	TR [msec]	Flip Angle	Partial Fourier	Pre-scan Normalise	GDC
3T Siemens Skyra (UK Biobank)	3D Axial	0.8mm × 0.8mm × 3mm	256x288x48	TE1=9.42 TE2=20	27	15°	87.5%	Yes	Off
3T Siemens Prisma FMRIB	3D Axial	0.8mm × 0.8mm × 3mm	256x288x48	TE1=9.42 TE2=19.7	27	15°	87.5%	Yes	Off
3T Siemens Prisma WH	3D Axial	0.72mm× 0.72mm× 3mm	256x288x48	TE1=9.42 TE2=19.7	27	15°	87.5%	Yes	Off
3T Siemens Trio JRH	3D Axial	0.8mm × 0.8mm × 3mm	280x320x48	TE1=9.42 TE2=16.7	27	15°	87.5%	Yes	Off
3T Philips Ingeia SPMIC	3D FFE Axial	0.8mm × 0.8mm × 3mm	288x288x48	TE1=9.4 TE2=20	27	15°	Off	Yes	On
3T Philips Achieva SPMIC	3D FFE Axial	0.8mm × 0.8mm × 3mm	288x288x48	TE1=9.4 TE2=20	27	15°	Off	Yes	On
3T GE MR750 QMC	3D Axial	0.8mm × 0.8mm × 3mm	288x288x46	TEs=9,13, 17,21,25, 30,34	42	15°	No but 90% undersampling in phase FOV	Yes	On

Diffusion MRI

The diffusion images were acquired with a monopolar pulsed gradient spin echo (PGSE) echo-planar imaging (EPI) sequence at $2mm$ isotropic spatial resolution. The phase encoding direction for all the scanners was in the anterior-posterior direction and blip-reversed spin-echo EPI images were acquired on all the scanner in order to generate fieldmaps to carry out geometric distortion correction for. Differences in gradient strength and simultaneous-multi-slice (multiband) acceleration capabilities affected the achievable minimum TE and TR across scanners. Both the Philips Achieva and GE MR750 missed multi-band capabilities, therefore the resulting TR was above 10 seconds. For the MR750, we opted for only relatively low b-value data (up to $b = 1000s/mm^2$), because of the low gradient strength and also the excessively long TR (TR was also long for the Philips Achieva, but the much stronger gradients allowed usable data in a reasonable scan time). This precluded the GE datasets from some of the summary features we extracted as they depended on multi-shell data (such as NODDI). Notice that in the absence of out-of-plane acceleration for the Achieva and MR750, in-plane parallel imaging with an acceleration of 2 was used to minimise TE. Angular resolution across b-shells was relatively constant across scanners. In summary, total scan times were in the order of 6.5 minutes for the Siemens scanners, 7.5 minutes for the Philips Ingenia, 18 minutes for the Philips Achieva and 12 minutes for the GE MR750.

Table 3.7: Protocol details for dMRI acquisition across the different scanners. The reference UK Biobank protocol is shown on the top row.

Scanner	Sequence	Voxel Size	Matrix Size	TE/TR	Parallel imaging	MB factor	Partial Fourier	b-values	Directions per b value	Number of b0's	Number of volumes	Total readout time [ms]
3T Siemens Skyra (UK Biobank)	Monopolar 2D PGSE EPI	$2mm^3$	104x104 x72	92/3600	No	3	75%	0,1000, 2000	50	5	105	69
3T Siemens Prisma FMRIB	Monopolar 2D PGSE EPI	$2mm^3$	104x104 x72	92/3600	No	3	75%	0,1000, 2000	50	5	105	69
3T Siemens Prisma WH	Monopolar 2D PGSE EPI	$2mm^3$	104x104 x72	92/3600	No	3	75%	0,1000, 2000	50	5	105	69
3T Siemens Trio JRH	Monopolar 2D PGSE EPI	$2mm^3$	104x104 x72	96.4/3600	No	3	75%	0,1000, 2000	50	5	105	69
3T Philips Ingenia SPMIC	Monopolar 2D PGSE EPI	$2mm^3$	112x112 x60	98/4293	1.5	3	80%	0,1000, 2000	50	5	105	74.3
3T Philips Achieva SPMIC	Monopolar 2D PGSE EPI	$2mm^3$	112x112 x56	70/10000	2	None	80%	0,1000, 2000	50	5	105	38.5
3T GE MR750 QMC	Monopolar 2D PGSE EPI	$2mm^3$	104x104 x56	72/11000	2	None	Off	0,1000	60	5	54	34.8

Resting-state functional MRI

The resting state functional MRI (rfMRI) images were acquired with 2D gradient echo planar imaging (GE EPI). All subjects were asked to keep their eyes open during scanning. Similarly as in dMRI, the difference in the MB capabilities of each scanner defined the minimum TR that could be achieved at a given spatial resolution. For Philips scanners, pushing the MB beyond 4 (it is 8 in Siemens scanners) caused excessive artefacts. We therefore opted for acquisitions that had the same spatial resolution to the Siemens scanners and roughly the same number of timepoints (400 in Philips vs 490 in Siemens), but differed in the temporal resolution and number of slices. For GE, we opted for the same spatial resolution, but this incurred a large penalty in the TR (no MB available), so we chose a roughly similar overall scan duration to the Philips scanners, keeping the number of timepoints to 200. In total the scan times were 6 minutes for Siemens scanners, 7.5 minutes for the Philips Achieva, 9.5 minutes for the Philips Ingenia and 7.5 minutes for the GE MR750. In each case, the flip angle was set to the Ernst angle for the corresponding TR, assuming T1=1500ms for grey matter at 3T.

Table 3.8: Protocol details for resting-state fMRI acquisition across the different scanners. The reference UK Biobank protocol is shown on the top row.

Scanner	Sequence	Voxel Size	Matrix Size	TE/TR	Parallel imaging	MB factor	Partial Fourier	Flip angle	EPI Factor	Number of volumes	Total readout time [ms]
3T Siemens Skyra (UK Biobank)	2D GE EPI	$2.4mm^3$	88x88 x64	39/735	No	8	No	52°	N/A	490	56
3T Siemens Prisma FMRIB	2D GE EPI	$2.4mm^3$	88x88 x64	39/735	No	8	No	52°	N/A	490	56
3T Siemens Prisma WH	2D GE EPI	$2.4mm^3$	88x88 x64	39/735	No	8	No	52°	N/A	490	56
3T Siemens Trio JRH	2D GE EPI	$2.4mm^3$	88x88 x64	39/735	No	8	No	52°	N/A	490	60
3T Philips Ingenia SPMIC	2D GE EPI	$2.4mm^3$	96x96 x48	39/1450	1.5	4	No	71°	55	400	39
3T Philips Achieva SPMIC	2D GE EPI	$2.4mm^3$	88x88 x64	39/1150	1	4	80%	65°	55	400	53
3T GE MR750 QMC	2D GE EPI	$2.4mm^3$	96x96 x50	39/2200	2	None	No	86°	N/A	200	25

3.3.5 Total Readout Time for EPI

Total readout time in EPI acquisitions is needed for performing fieldmap-based corrections of susceptibility-induced distortions (Andersson et al. 2003). In addition to the above modalities, spin-echo EPI data were collected with reversed phase-encode blips, resulting in pairs of images with susceptibility-induced distortions going in opposite directions. From these pairs, the susceptibility-induced off-resonance field can be estimated using the method described in (Andersson et al. 2003) as implemented in FSL's topup (Jenkinson et al. 2012). The total readout time for the fieldmap and the acquisition are required in order to estimate and apply this off-resonance field to dMRI and fMRI data and perform correction of susceptibility-induced distortions. In the simplest case of non-accelerated 2D acquisitions, this would be trivial to obtain, but the calculation is complicated in the presence of in-plane accelerations and on what acquisition properties are exactly reported by each vendor in certain DICOM header tags.

FSL's topup needs the "effective" total readout time to perform the correct calculations. This is now automatically calculated by dcm2niix and saved in json files but it was not at the time of the work hence requring the follwing calculation.

As defined by the Brain Imaging Data Structure (BIDS) format (Gorgolewski et al. 2016), the "effective" total read out time is defined as the readout duration that would have generated the data with the given level of distortion. This is linked to the "effective" echo spacing, which is the sampling interval between lines in the phase-encoding direction based on the size of the reconstructed image in the phase-encoding (PE) direction (i.e. taking into account in-plane accelerations). We computed these terms using a formula given in (Rorden et al. 2012) which was consistent for all vendors:

$$TotalReadoutTime_{eff} = EchoSpacing_{eff} \times (ReconMatrix_{PE} - 1) \quad (3.1)$$

where the effective echo spacing is provided by dcm2niix (Li et al. 2016) and takes into account corrections for in-plane accelerations.

3.3.6 Data Quality Control

To characterise the level of consistency in data quality across sites and scanners, we performed quality control (QC) using established frameworks: MRIQC (Esteban et al. 2017) (for T1w and fMRI) and EDDYQC (Bastiani et al. 2018) (for dMRI). MRIQC calculates a number of metrics and subsequently uses a supervised machine learning framework to classify data as either acceptable or unacceptable. EDDYQC uses the outputs of FSL's comprehensive distortion and motion correction package (EDDY) (Andersson & Sotiropoulos 2016) to extract features that characterise different aspects of dMRI data quality.

A number (35) of Image Quality Metrics (IQMs) calculated by these tools are summarised in Figure 3.2. Overall, the IQMs can be split into distinct categories that characterise various aspects of data quality: from noise and signal levels, to motion, distortions and artefacts. Below, we provide definitions for each of these metrics.

Noise-Level Measures

The **Signal to Noise Ratio (SNR)** is amongst the most standard metrics quantifying the level of true signal with respect to the level of noise. The measure of SNR proposed by Dietrich et al. (Dietrich et al. 2007*a*) is used here for T1w images, where the background of the image is used to estimate the noise variance (i.e. assuming that no true signal exists in the background and all signal there is pure noise). Signal is obtained as the average intensity

QC Metric	Category	T1	dMRI	rfMRI
Spatial Contrast to Noise Ratio	Noise related	✓		
Angular Contrast to Noise Ratio			✓	
Signal to Noise Ratio		✓	✓	
Temporal SNR				✓
Coefficient of Joint Variation (CJV)	Intensity related	✓		
AFNI's Outlier index				✓
AFNI's Outlier Ratio				✓
Outliers (Intensity dropout)			✓	
Quality Index 1		✓		
Absolute Motion	Motion related		✓	✓
Relative Motion			✓	
Eddy Current Distortions	Distortion related		✓	
Susceptibility Induced Distortions			✓	
Full-width-half-maximum	Blurring related	✓		✓
Entropy Focused Criterion (EFC)		✓		

Figure 3.2: *Summary of IQMs for different categories and the imaging modalities they are applicable to.*

from a homogeneous region containing data (e.g. white or grey matter). An alternative measure of SNR uses repeats of the same scan (when such repeats exist, such as in rfMRI or the b=0 volumes in dMRI). In that case, the average of signal intensities across repeats quantify the signal, while the variance across repeats qunatifies the noise. SNR for dMRI in b=0 images is calculated that way.

$$\text{SNRd} = \frac{\langle S_{\text{Tissue}} \rangle}{\sigma_{\text{Background}}} \tag{3.2}$$

and

$$\text{SNR}_{\text{repeats}} = \frac{\langle S_{\text{Tissue}} \rangle}{\sigma_{\text{Tissue}}}, \tag{3.3}$$

where $\langle S \rangle$ is the average signal across repeats. The **temporal SNR (tSNR)** (Krüger & Glover 2001) is used to evaluate data quality for fMRI images, and is based on the $SNR_{repeats}$ equation, where the number of repeats are as many as the fMRI timepoints.

The **contrast-to-noise-ratio (CNR)** (Magnotta & Friedman 2006) is an important measure of scanner performance as it quantifies contrast.For anatomical images, the contrast is calculated in the spatial domain depicted by the difference in the mean intensities between different tissue types, for instance white matter and grey matter, while the noise can be depicted as the variance in the intensities of the background. It is calculated in the following way:

$$\frac{\langle S_{\mathrm{GM}} \rangle - \langle S_{\mathrm{WM}} \rangle}{\sigma_{\mathrm{Background}}} \tag{3.4}$$

For dMRI, a different version of CNR is used that quantifies contrast in the angular (diffusion) domain (Bastiani et al. 2018). This **angular CNR** quantifies for each b-shell the difference between the min and max dMRI signal intensity over the noise in the dMRI data.

Intensity-related Measures and Outliers

The **Coefficient of Joint Variation (CJV)** (Ganzetti et al. 2016) captures joint Variation between signal intensities of white matter and grey matter in T1w images. It is calculated the following way :

$$\mathrm{CV}_{\mathrm{WM}} = \frac{\sigma(\mathrm{WM})}{\mu(\mathrm{WM})}, \mathrm{CV}_{\mathrm{GM}} = \frac{\sigma(\mathrm{GM})}{\mu(\mathrm{GM})}, \mathrm{CJV} = \frac{\sigma(\mathrm{GM}) + \sigma(\mathrm{WM})}{\mu(\mathrm{WM}) - \mu(\mathrm{WM})} \tag{3.5}$$

where σ and μ are the standard deviation and the mean intensity of the given tissue class respectively. Higher values indicate heavier head motion and/or the presence of large spatial biases or intensity inhomogeneities.

The Analysis of Functional NeuroImages (AFNI) (Cox 1996, Cox & Hyde 1997) software is used to compute outliers for fMRI data. **AFNI's outlier ratio** gives the mean fraction of outliers per fMRI volume based on signal intensity. **AFNI's quality index** is a mean quality index across the time series. Small values of the index are 'good', indicating that a particular volume is not very

different from the norm.

dMRI Outliers summarises the percentage of slices classified as outliers per dMRI volume and across the whole dataset. Outlier slices are identified by EDDYQC as those suffering signal dropouts which is usually caused by subject motion. They are grouped by b-value (for our data either b_{1000} or b_{2000}).

The **first Quality Index (QI_1)** proposed in (Mortamet et al. 2009) gives the proportion of voxels with intensities corrupted by artefacts with respect to the number of voxels in the background. //

Motion-related Measures

Absolute and Relative motion summarise subject motion between volumes in dMRI, consisting of three translations and three rotations across the x,y and z axes. Absolute motion for each volume is calculated with respect to a single reference volume (e.g. first b=0), whereas relative motion is calculated with respect to the previous volume.

Similar to these definitions, MRIQC outputs the **Framewise Displacement (FD)** metric for fMRI data. This is the average motion summarised throughout the duration of the acquisition. It most closely resembles the absolute motion from dMRI.

Artefacts and Distortions-related Measures

Eddy current-induced distortions are quantified by EDDYQC as the variability in these distortions that cause misalignment from volume to volume across dMRI scans.

Susceptibility-induced distortions caused by off-resonance fields due to differences in magnetic susceptibility at tissue-air interfaces are quantified by EDDQYQC for dMRI data.

Blurring related Measures

The **Full-width Half-maximum (FWHM)** (Friedman et al. 2008) gives an estimation of how blurry the image is.

Finally, the **Entropy Focused Criterion (EFC)** (Atkinson et al. 1997) gives and indication of ghosting and blurring in the image caused by head motion. It can be applied for both 3D and 4D images and is returned by MRIQC for both T1w and fMRI data.

3.3.7 IQM Comparisons within scanners to Assess Consistency of scan-rescan

We used the IQMs derived from dMRI to compare the within scan sessions. These IQMs provide measures such as subject motion and outliers which will allows us to directly quantify how stable each of the within-scan repeats were.

3.3.8 IQM Comparisons Between Scanners to Assess Consistency of Image Quality

For each scanning session, IQMs were calculated which produced summary measures for anatomical (T1-weighted), microstructural (dMRI) and functional (rfMRI) data. A correlation matrix was constructed which showed the correlation of IQMs across sessions. This was performed using the data from the 4 subjects for whom repeat scans were acquired to allow the comparison to be made between the consistency of IQMs across different scanners and within the same scanner. For each of the 4 subjects, the 5 between-scanner repeats were merged with the 6 within-scanner repeats to give a matrix A with dimensions 11 (number of scanning sessions) $\times$ 35 IQMs. Each IQM column of data was z-scored then each row of this matrix was normalised with the magnitude as to make each row a unit vector. The matrix product of $A \times A'$ gave an

11×11 correlation matrix where each element was a correlation of IQMs from two scanning sessions. The final correlation matrix is displayed in Figure 3.6.

3.4 Results

3.4.1 Examples of Collected Data

In total, 80 scanning sessions were performed (in addition to piloting ones) to collect multi-modal brain MRI data from 10 subjects in 6 scanners (between-scanner sessions), while 4 subjects underwent an extra 5 within-scanner repeats. In this section we provide representative examples of the acquired data.

Figures 3.3 summarises examples from all between-scanner multi-modal data for a single subject. We can qualitatively observe relatively decent quality and contrast for all modalities in all scanners, although as expected there are appreciable differences. These between-scanner differences can vary with the type of the imaging modality. Figures 3.4 and 3.5 show examples of modalities where this between-scanner difference is appreciably high or less noticeable respectively. Figure 3.4 illustrates how the FA maps from the diffusion data differ appreciably across scanners, while when the same subjects is scanned in the same scanner the variability is much less. Specifically, there is a noticeable difference in contrast between grey matter and white matter (grey matter regions indicated by yellow arrows). Furthermore, there is less noise and spatial inhomogeneity in intensities in some scanners compared to others (example of this in regions indicated by red arrows) and this variability is less in the within-scanner repeats. On the other hand, Figure 3.5 shows that for T1-weighted images, the between and within-scanner variabilities are much more comparable. These results provide an early qualitative demonstration that inter-site effects and the need for harmonisation are not necessarily equivalent across imaging modalities and features.

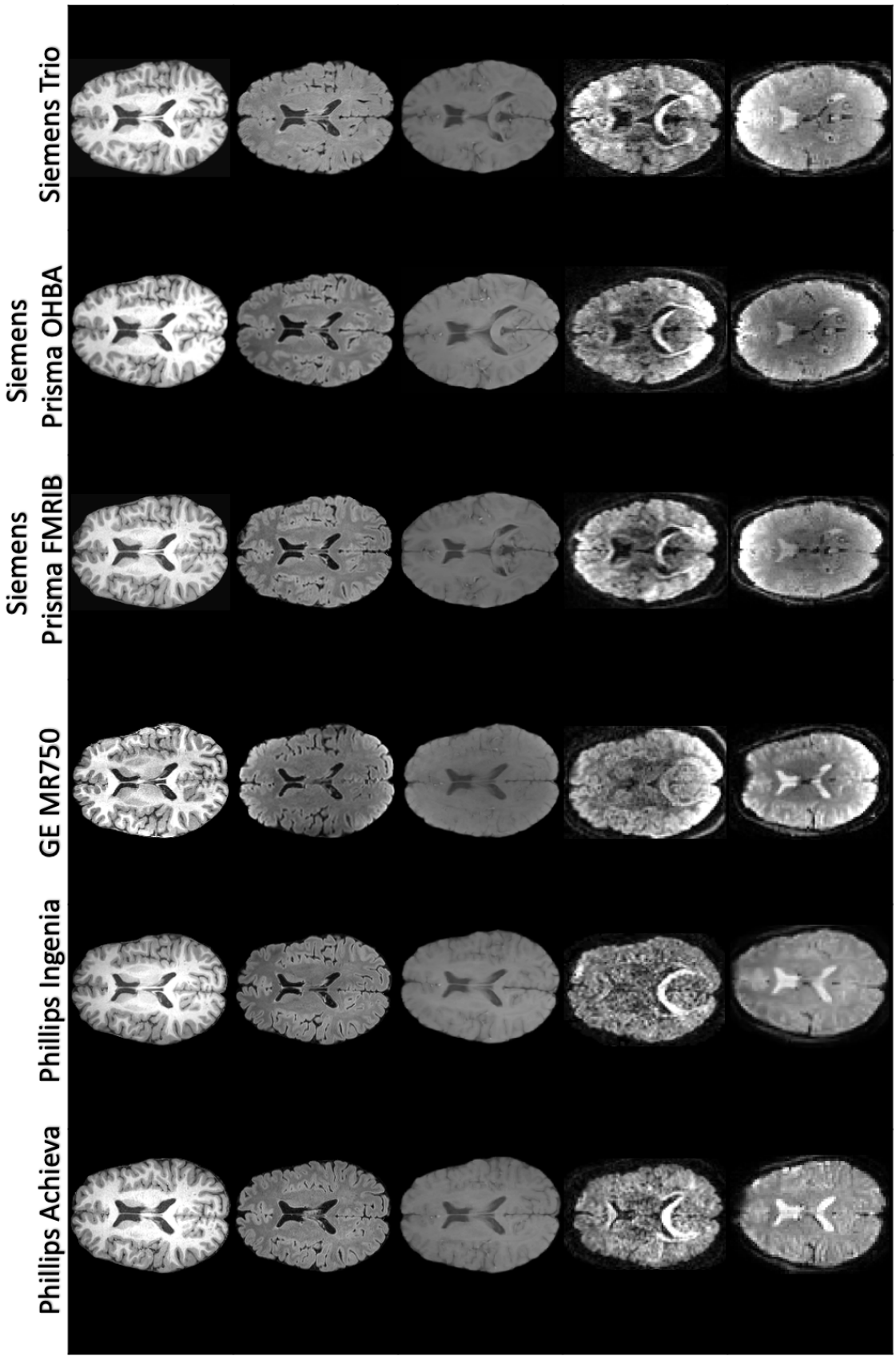

Figure 3.3: *Illustration of all acquired multi-modal data for a single subject across all 6 scanners and 5 imaging modalities. SWI image is a magnitude image and dMRI image from a* $b = 1000s/mm^2$ *shell*

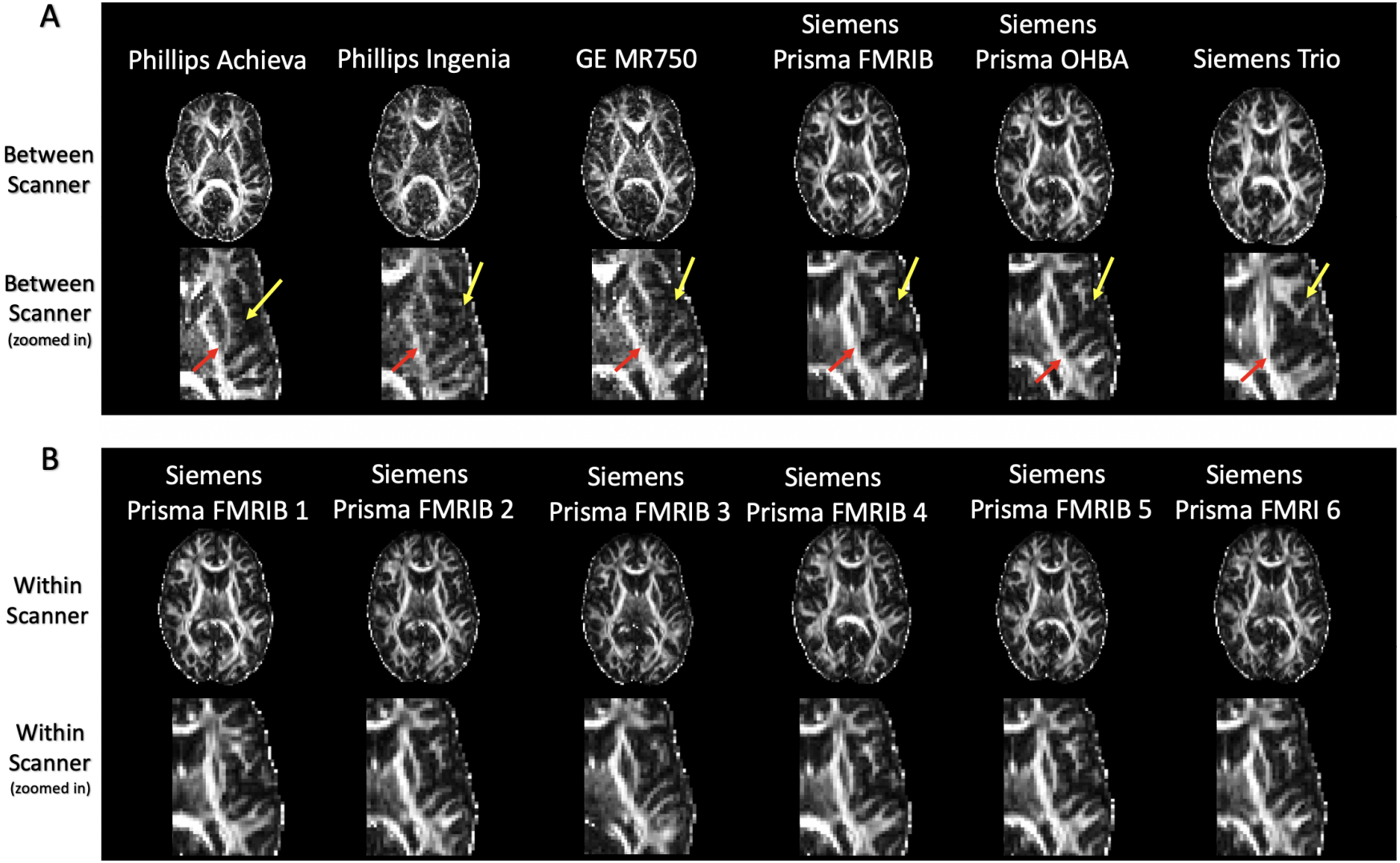

Figure 3.4: *A comparison of FA maps from acquired dMRI images between (A) different scanners and (B) within the same scanner. There is a noticeable difference in contrast between grey matter and white matter (grey matter regions indicated by yellow arrows). There is also less noise and spatial inhomogeneity in intensities in some scanners compared to others (example of this in regions indicated by red arrows) and this variability is less in the within-scanner repeats.*

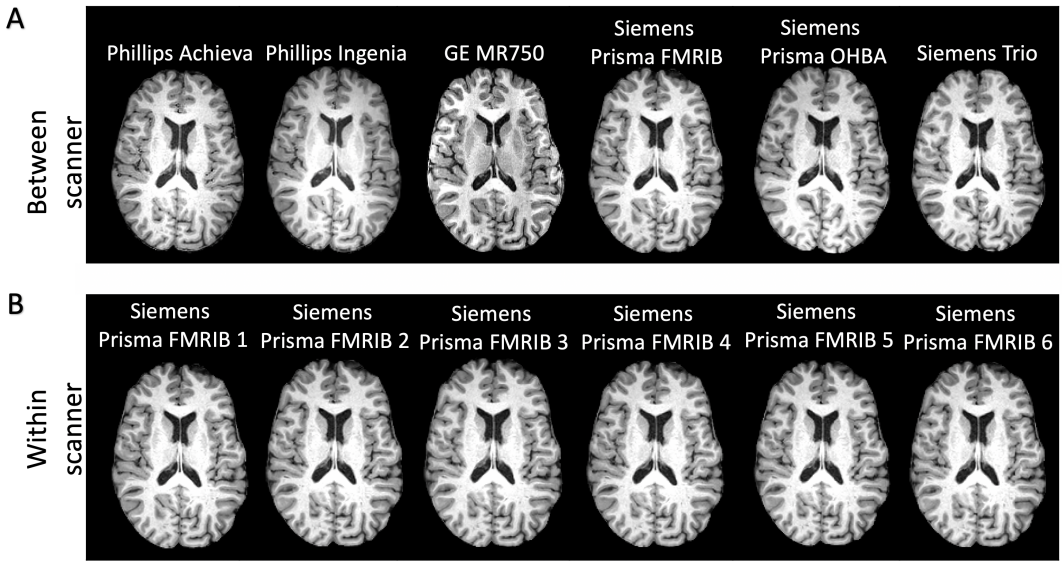

Figure 3.5: *A comparison of acquired T1-weighted images between (A) different scanners and (B) within the same scanner. There are few noticeable qualitative differences in the data between scanners. This is comparable to within-scanner data which is similarly consistent. Intensities of images have been scaled between the* -90^{th} *and* 90^{th} *percentile.*

3.4.2 QC and Data Quality Comparisons

Quality control was performed for all the scanning sessions, as described in methods. For each scanning session, IQMs were calculated that reflected data quality for anatomical (T1w), microstructural (dMRI) and functional (rfMRI) data. As expected, IQMs were more variable across the six between-scanner repeats rather than the six within-scanner repeats for the same subject. Figure 3.6 demonstrates for a set of subjects (those with within-scanner repeats) the correlation of IQMs between all scanning sessions. As we can see, the within scanner repeats are more consistent quality wise compared to the between-scanner sessions as shown by the larger correlation values in the region of the matrix representing within-scanner repeats.

We further used IQMs to assess the consistency of image quality across *scanners*. To do that, each quality metric for each subject data was z-scored across the 6 scanners. The Z-scores were then averaged across the 10 subjects. A colour-coded map of the z-scored IQMs is shown in Figure 3.8 for each IQM

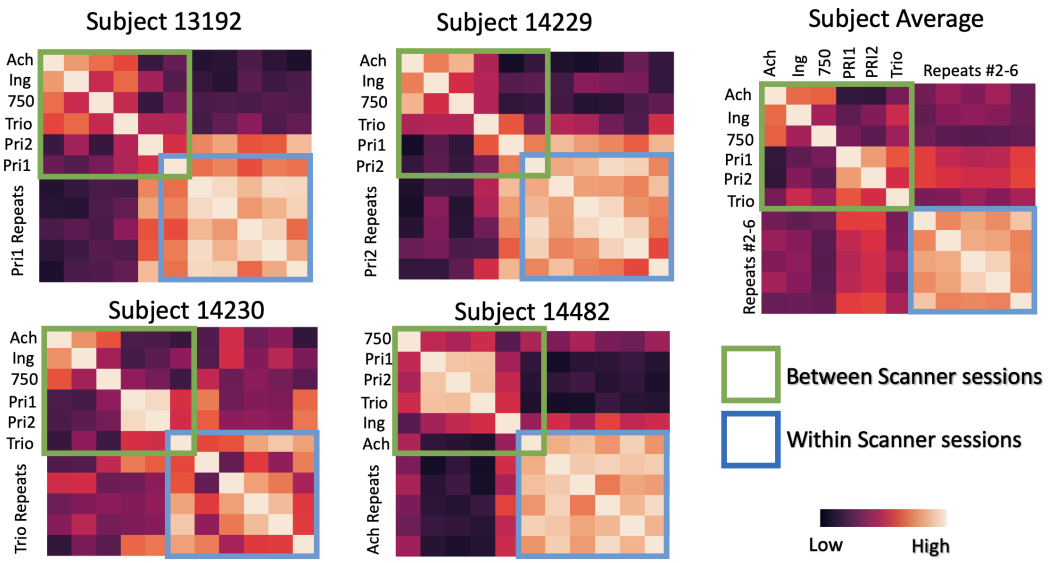

Figure 3.6: *Correlation matrices depicting correlation of IQMs between scanning sessions for subjects with repeat scans. On average, within-scanner repeats are more consistent quality wise than between-scanner sessions. The range of the colour bar is from the -90^{th} to 90^{th} percentile of values. Outliers and CNR for $b = 2000s/mm^2$ have been excluded.*

and scanner, showing consistency of average image quality across scanners for the different modalities and metrics. We can observe that all metrics for all scanners are within 2 standard deviations of their respective means, i.e. there are no major outliers in terms of raw image quality and/or artefacts (73% of the IQMs are within one standard deviation from their respective means). The 3 Siemens scanners seems to be closer overall to the means (i.e. z scores closer to zero), but there are modality-specific differences. The Philips Achieva T1-weighted and dMRI data are also closer to the mean scanner quality, while the GE rfMRI is closer to the respective rfMRI IQM means.

Similarly to the above analysis, we used the IQMs to assess the consistency in image quality across *subjects*. Each quality metric for each scanner was z-scored across the 10 subjects. The Z-scores were then averaged across the 6 scanners. A colour-coded map of the z-scored IQMs is shown in Figure 3.9 for each IQM and subject. We can observe that the vast majority of IQMs (93%) are within one standard deviation from their respective means. This provides evidence that there were no outlier subjects in our cohort.

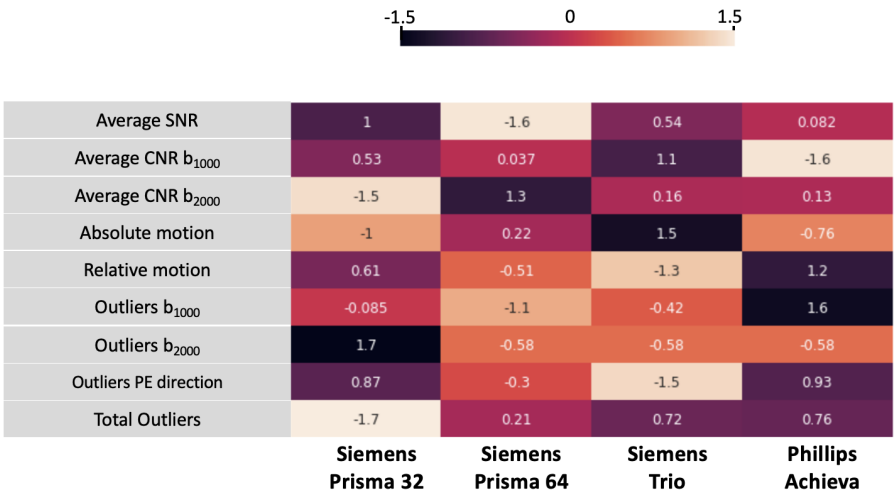

Figure 3.7: *Heatmap of Image quality metrics (IQM) across the 6 scanners. Z-scores were taken of average IQM values across the 6 within scan repeats. Higher positive or negative values represent large deviations from the mean.*

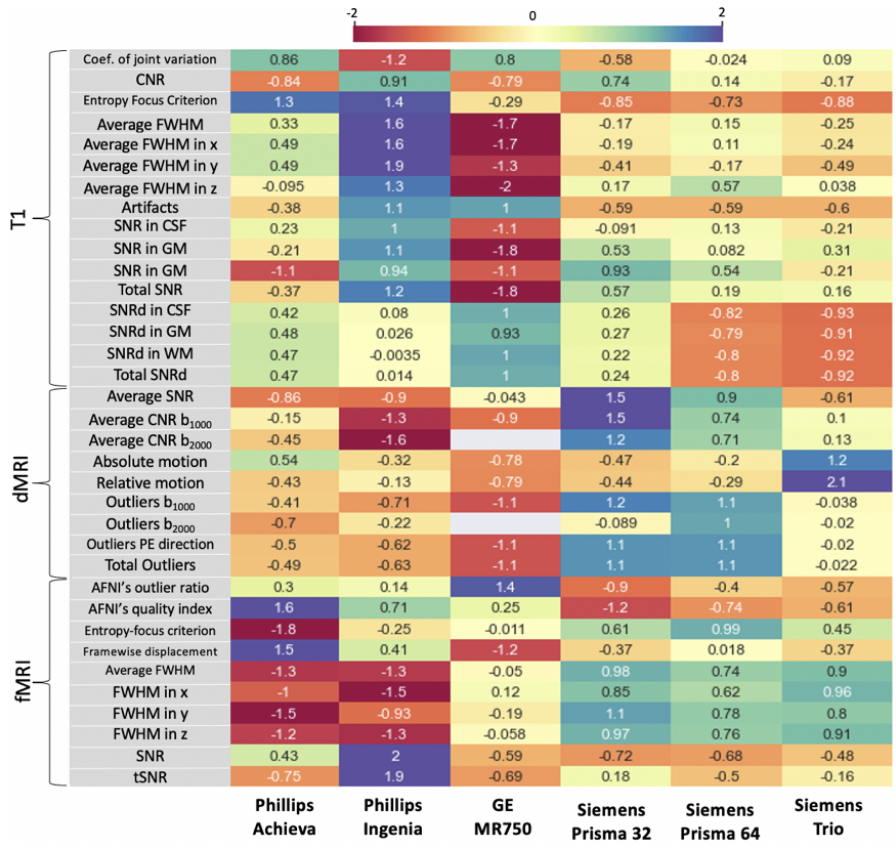

Figure 3.8: *Heatmap of Image quality metrics (IQM) across the 6 scanners.Each quality metric for each subject was z-scored across the 6 scanners. The Z-scores were then averaged across the 10 subjects. Higher positive or negative values represent large deviations from the mean. *Single-shell data was acquired on the GE scanner, hence the absence of $b = 2000s/mm^2$ dMRI quality metrics.*

Figure 3.7 shows the degree to which the within-scan repeats were consistent.

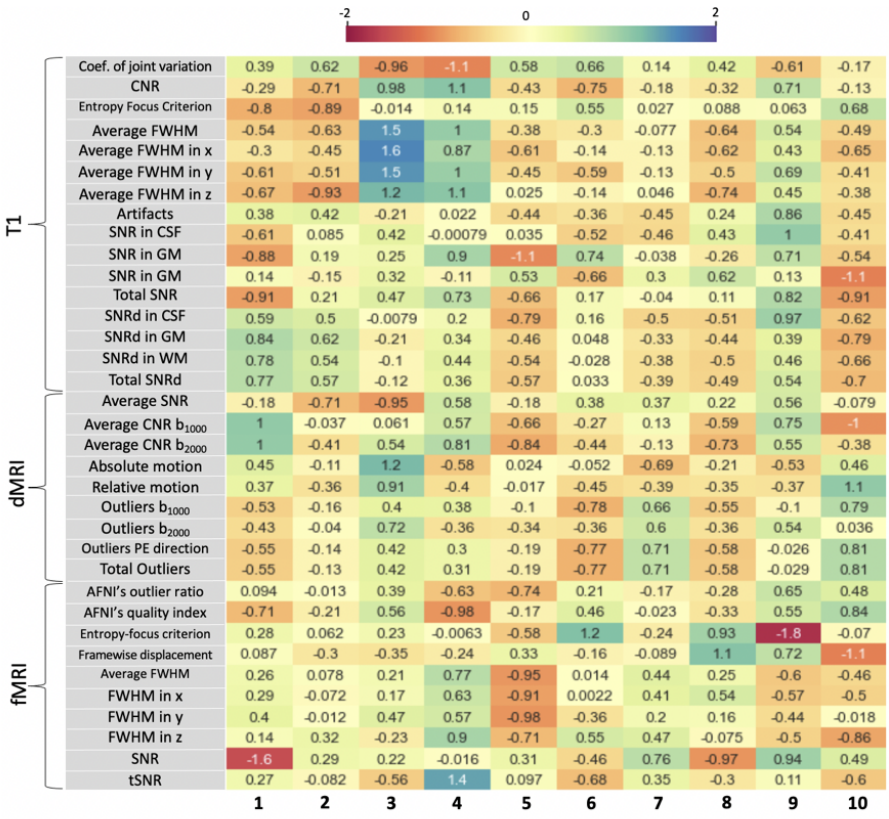

Figure 3.9: *Heatmap of Image quality metrics (IQM) across the 10 subjects. Each quality metric for each subject was z-scored across the 10 subjects. The Z-scores were then averaged across the 6 scanners. Higher positive or negative values represent large deviations from the mean. *Single-shell data was acquired on the GE scanner so those have been excluded from the averaging of the $b = 2000s/mm^2$ dMRI quality metrics.*

The within-scanner repeats obtained from the Philips Achieva were affected most by relative motion whereas the the repeats obtained from the Siemens Trio were affected the most by absolute motion. The high degree of motion associated with the Philips Achieva is likely due to a 4 of these scans being performed back to back in pairs of 2 (2 scans back to back on one day and an additional 2 on another day). Prolonged scan times have been known to contribute to an increase in subject motion (Zaitsev et al. 2015).

3.4.3 QC Across Scanners and Subjects

Figures 3.8 and 3.9 shows that while there is some variability in quality metrics across scanners and subjects, there are no extreme outliers. This suggests that the image quality across all scanning sessions was consistent.

Though there are no extreme outliers, the average FWHM and SNR values for T1 weighted imaging are lower for the GE MR750 than for the other scanners. A major difference between the GE scanner and the other scanners is that that the GE runs a BRAVO sequence for the T1-weighted imaging whereas the Philips and Siemens scanners run MPRAGE sequences. There is evidence in literature in support of this as it has been stated in (Bugge et al. 2017) that MPRAGE sequences were initially restricted to Siemens scanners but have since then been made available on other vendors.

The IQM trends also indicate a discrepancy between the two methods of determining SNR. The SNR values which use the air background of a reference (SNRd) suggest that images from the GE scanner have the largest SNR. In contrast, the SNR values calculated using the standard deviation of tissue as noise (SNR), suggest that images from the GE scanner have the lowest SNR. This is pointed out in (Dietrich et al. 2007*b*) where it is stated that SNR values calculated using two regions may not agree with the true SNR, as noise in modern acquisitions varies spatially and therefore using the background to get noise variance can be suboptimal. Therefore the SNR values calculated using the standard deviation of tissue as noise are a more appropriate measure for this scope.

Furthermore, it can be observed from the fMRI quality metrics that there is a positive correlation between the FWHM in fMRI and tSNR. This is consistent with previous studies (Molloy et al. 2014) which have shown that spatial smoothing improves temporal SNR.

3.4.4 Within-vendor QC Differences

We used the IQMs to perform additional comparisons and characterise within-vendor quality differences. The range of systems and hardware we have considered in our study allows for a number of comparisons, such as i) the effect of maximum gradient strength in Siemens Scanners, ii) the effect of bore size in Philips scanners, iii) the effect of using a 32 vs a 64-channel in Siemens Prisma scanners. The IQM's Figure 3.10 were chosen because these are known to be affected by the hardware features we are considering. For example, it has been reported that a higher number of coil channels results in higher SNR (Keil et al. 2013).

Maximum gradient strength

We compared a selection of IQMs between two scanners from the same vendor, but with different maximum gradient strength. The Siemens Prisma (FMRIB) has $80mT/m$ gradients, while the Siemens Trio has $45mT/m$. As shown in Figure 3.10A, the SNR benefits are evident for higher maximum gradient strength. For T1 and dMRI modalities, there is a gain in total SNR for scanners with higher gradients. The angular CNR for dMRI and tSNR for rfMRI are also higher for the system with higher gradients. This confirms the trend reported (Hidalgo-Tobon 2010) where it is stated that high gradient strengths are beneficial to image quality. Interestingly, dMRI motion is on average less for the high gradient system. This may seem counter-intuitive, as higher gradients lead to more shaky acquisitions. But it may reflects the fact that newer generation scanners (such as the Prisma) actively compensate for the scanner table motion to improve patient comfort.

Bore size/Maximum gradient strength

To probe the effect of bore size on image quality, we looked into the two Philips scanners with different magnet bore sizes (which however also had different gradient systems): the wide-bore Philips Ingenia ($70cm$ bore, $45mT/m$

maximum gradients) and the narrow-bore Achieva ($60cm$ bore, $80mT/m$ maximum gradients). Larger bore sizes in principle contribute to patient comfort (Oztek et al. 2020) and potentially less motion. The comparison is shown in Figure 3.10B and confirms this expectation, by showing less absolute motion for dMRI and framewise displacement for rfMRI for the scanner with a wider bore. For dMRI, the angular CNR is greater for the system with a narrower bore and a higher gradient which is also as expected.

A slightly unexpected result (Liney et al. 2013) is depicted by the higher T1 SNR for the wide-bore scanner compared to the narrow-bore system. This can be explained in part by the higher average FWHM of the wide-bore scanner than the narrow-bore scanner as shown in Figure 3.8. The average FWHM z-score value for the wide bore Ingenia scanner is 1.6 whereas for the Achieva it is 0.33. This indicates a higher degree of blurring for the wide-bore system and therefore causes it to have an "artificially" higher SNR. An additional contributing factor may be that the Ingenia is a newer system than the Achieva and due to better patient comfort there are less intensity outliers/overall better quality index. Evidence of this is shown in Figure 3.8 (CJV z-score of -1.2 for the Ingenia and 0.86 for the Achieva) meaning less intensity related outliers. Combined, these factors explain why, eventhough the Ingenia is a wide-bore system, it outperforms the Acheiva in terms of SNR.

Number of coil channels

The comparison between the Siemens Prisma 1 (FMRIB) and the Siemens Prisma 2 (OHBA) in Figure 3.10 C demonstrates the effect of a 32 channel head coil compared to a 64 channel head coil, when effectively everything else is the same. An increased number of coil channels has been reported to provide an increase in SNR, particularly at the cortex compared to deeper structures in the brain (Keil et al. 2013). What can be clearly seen in Figure 3.10 is that, for all imaging modalities, the opposite is true. This suggests that the 64-ch Siemens coil may not be implemented to its full potential. For T1-weighted

images, we report both the SNR in white matter and in grey matter. We see that there is a decreased difference in SNR between the two coils in grey matter compared to white matter.

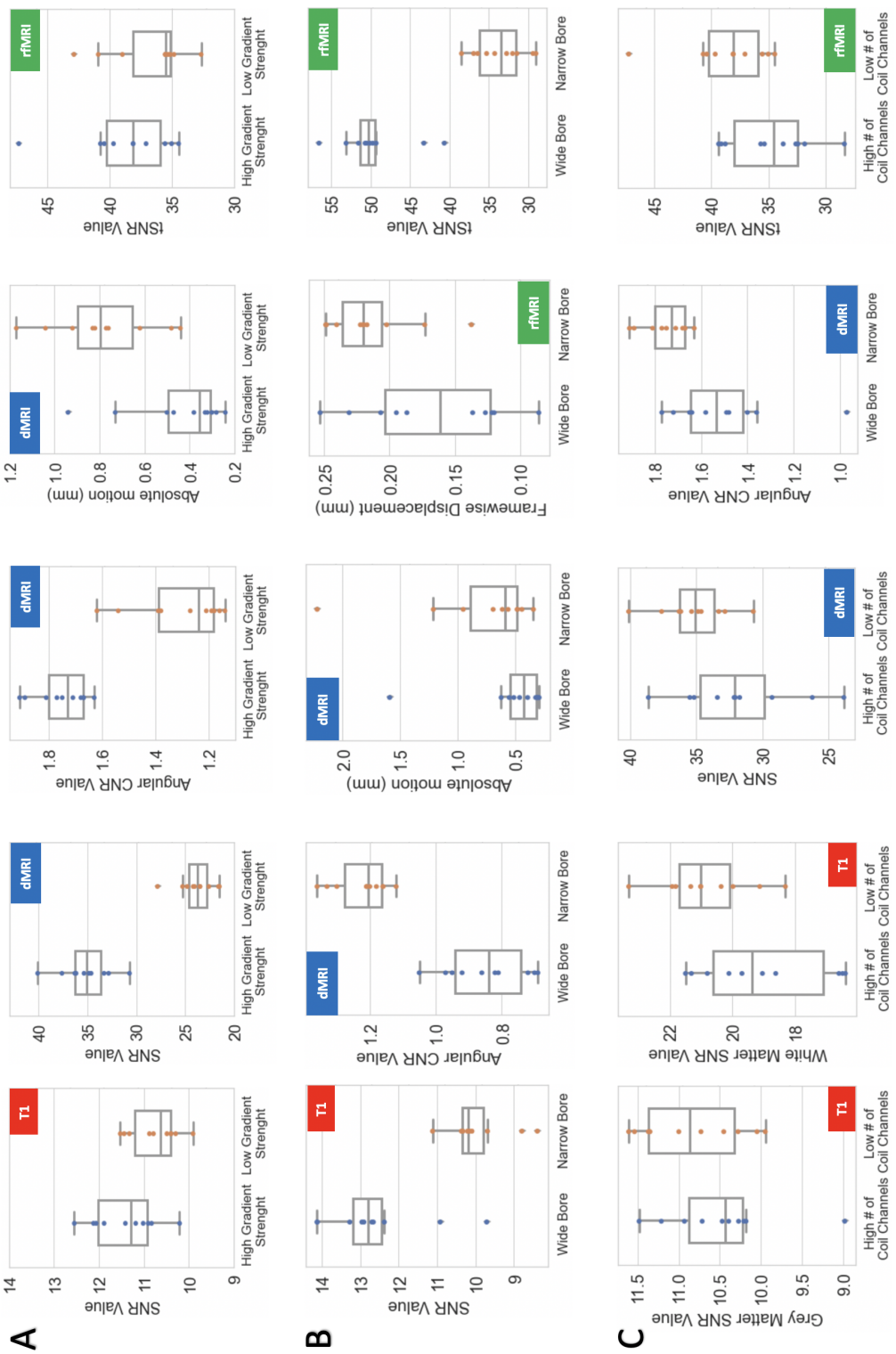

Figure 3.10: *An exploration of within-vendor QC differences due to specific hardware features. Boxplots are depicting the distributions across ten different subjects. A) Max gradient strength (Siemens Prisma (80mT/m) vs Siemens Trio (45mT/m)) B) Bore Size and max gradient strength (Philips Ingenia vs Philips Achieva) C) Number of Coil Channels(Siemens Prisma - 64ch vs Siemens Prisma - 32 ch)*

3.5 Discussion

We have presented the acquisition setup of a novel harmonisation resource for multi-modal neuroimaging data, based on a travelling-heads paradigm. We have characterised data quality across the 80 scanning sessions and identified interesting trends in the data. We have quantitatively presented the image quality by comparing the variability in image quality between different scanners with the image quality within scanners. To compliment this, we have qualitatively shown appreciable between-scanner variability in certain imaging modalities when compared to within-scanner variability of the same modality for the same subject.

Besides ensuring consistent data quality, we also used derived image quality metrics to demonstrate how scanner hardware differences quantitatively affect data quality such as the number of channels in a head coil, the bore size or the maximum gradient strength of a scanner. For several cases, the results reflected what is to be expected. Interestingly, there were some cases where other factors such as having a more modern scanner outweighed the effects of hardware differences alone.

The QC we performed was limited in that we did not extract metrics for T2 FLAIR and SWI. All of the IQMs extracted from the T1w images could have also been extracted from T2 images (Esteban, Blair, Nielson, Varada, Marrett, Thomas, Poldrack & Gorgolewski 2019), which wasn't readily apparent at the time of processing. In theory, this could also be performed on the SWI images. The IQM's suitable for this would be CNR and SNR as demonstrated in (Borrelli et al. 2015).

Compared to previous travelling-head studies (Tax et al. 2019, Kurokawa et al. 2021, Hawco et al. 2018, Duff et al. 2021), our study extends them in a number of ways: i) data is acquired from all 3 major vendors and from different

generations of scanners from the same vendor, ii) data is acquired from physically different imaging sites where radiographers and practices are different, iii) data is acquired from many neuroimaging modalities, iv) scan-rescan data is acquired which allows the assessment of within-scanner, within-subject variability in addition to between-scanner variability.

As mentioned previously, the data acquisition was interrupted by the Covid-19 lockdowns, which resulted to longer than intended time intervals between the repeated scans of some subjects (range of 6 - 22 months, median interval was 15 months). This is longer than the the time interval between scans from other studies, but not significantly longer taking into account two lockdowns: (Tax et al. 2019) Average time: 21-22 months, (Kurokawa et al. 2021) Range: 1-5 months , (Hawco et al. 2018) Range: 1-36 months , (Duff et al. 2021) Range: 1-14 days. The relatively large time interval between scans in our study could introduce nuisance effects due to ageing, however these are expected to be small due to the young age of the cohort and are not expected to drive results. Additionally, we do not have QC metrics for SWI as there is no standardised framework for obtaining SWI image quality metrics. A further limitation is that there are certain discrepancies in how the data were acquired across the scanners. Specifically, multi-shell diffusion data was not acquired for all the scanners, neither did all the scanners have multi-band capabilities. Adjustments were made to account for this yet in some cases, these discrepancies had clear consequences, namely the omission of NODDI processing from the GE diffusion data. Nevertheless, these differences represent a realistic scenario.

The dataset and results presented in this chapter are a foundation to the material of subsequent chapters. We will explore the vast amount of features that can be derived per subject as a result of the wide range of imaging modalities acquired.

3.6 Summary

This chapter introduced the harmonisation dataset acquired. The value of this data has been demonstrated by comparing it with current datasets in the literature and it has been shown that this data is more comprehensive and wields greater potential for both the assessment and development of harmonisation techniques and algorithms. This chapter also gave an overview of the protocols used in the brain imaging component of the UK Biobank Pipeline and some of the compromises that were made when matching the imaging protocols of the 6 scanners used in this project to those originally used in the UK Biobank. We have also shown how consistent the imaging data is across the scanners and subjects. The dataset has been used to assess how within vendor hardware differences influence image quality showing that the utility of this dataset extends beyond the field of harmonisation.

This dataset will be used in Chapter 4 to map between-scanner variability for a large number of imaging-derived features. The within scanner repeats will be used as a baseline for comparison which will allow assessment of between scanner variability with respect to repeated measurements of the same subject in the same scanner. Imaging data from multiple subjects from the UK Biobank brain imaging study will also be used as a baseline for comparison, which will allow the assessment of between-scanner variability with respect to biological variability. The multi-modal nature of the data will allow demonstration of which modalities and imaging features seem to be more and less robust against inter-site/inter-scanner effects.

Finally, in Chapter 5, the the potential of this dataset will be further displayed as a testbed to evaluate existing harmonisation approaches, both implicit and explicit ones. This will be done by identifying optimal pipelines and processing steps that minimise between-scanner variability and also by testing

post-processing harmonisation tools.

3.7 Appendix: Modifications to Philips Achieva dMRI Data

All $b = 0\ s/mm^2$ volumes during the dMRI acquisitions were interspersed within the full diffusion protocol for all scanners. This is a slightly more optimal approach for distortion and motion correction compared to the alternative of having all b=0 volumes acquired at the beginning, as it avoids differences in motion susceptibility (and relevant downstream effects) between $b = 0\ s/mm^2$ and DWI images (subjects typically move less at the beginning of a scan). However, a scanner software glitch on the Philips Achieva did not allow expected performance when we tried to intersperse the b=0 images in the full q-space sampling protocol, even if in theory it allowed this interspersion to run and returned data. In particular, the interspersing caused ghosting for the intermediate $b = 0$ compared to the intensity of the first $b = 0$. (see Figure 3.11).

To resolve the issue we removed the interspersed $b = 0$ volumes and instead modified the Achieva dMRI protocol to acquire extra $b = 0$ volumes independently, before and after the main dMRI protocol (3 just before and 3 just after) and merged these with the rest of the dMRI data. To alleviate potential baseline intensity differences between the independent b0s and the full dMRI data, we matched all bandwidth and timings as close as possible to the full dMRI protocol ones. This still left some baseline intensity differences, which we removed by normalising the intensities of the 6 independent b=0 volumes, $I_{k=1,2,3,4,5,6}$,to match the intensity of the $b = 0\ s/mm^2$ in the full acquisition I_{orig}. Specifically, we multiplied each independently acquired $b = 0$ volume by the the ratio $\frac{I_{orig}}{I_{\mathbf{k}}}$. Figure 3.11B shows the time series data before and after intensity normalisation.

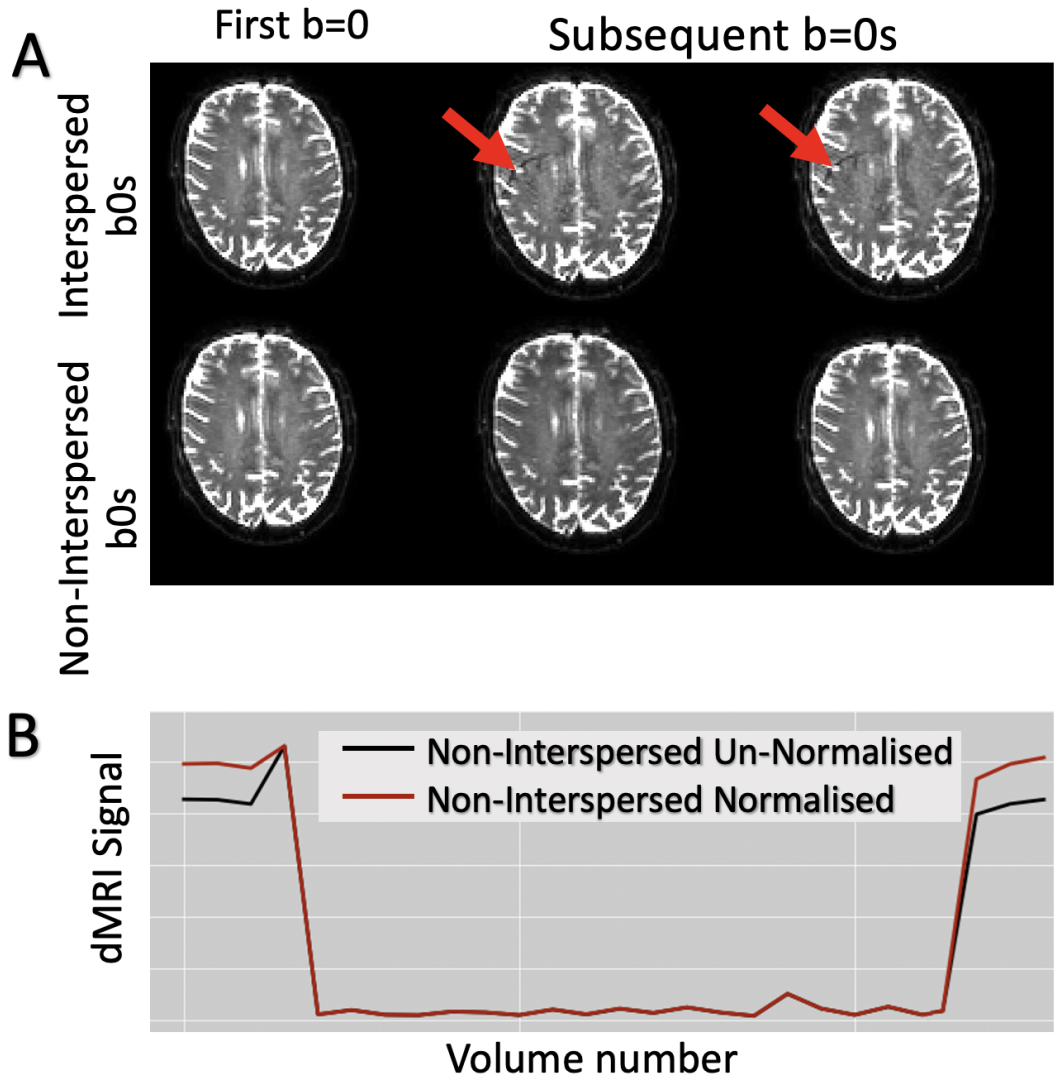

Figure 3.11: *Resolution of artefacts in dMRI data from Philips Achieva. A) The ghosting caused by having interspersed $b = 0$ voloumes. B) The intensity difference between the 3 independently acquired $b = 0$ volumes before the main protocol and the first $b = 0$ volume of the main protocol after and before normalisation.*

Chapter 4

Mapping Inter-scanner Variability for Multi-modal Imaging-derived Features

In the previous chapter, we presented the acquisition setup of a novel harmonisation resource for multi-modal neuroimaging data. We showed the various ways in which it is more comprehensive than currently existing data sets. These include the wide range of scanner vendors used, the variety of imaging modalities, and the travelling heads paradigm utilised. In this chapter, we use this data to explore how inter-scanner variability is reflected to a large set of imaging-derived features, such as sub-cortical structure volumes, micro structural measures and connectivity metrics. Collectively such features are referred to as Imaging-Derived Phenotypes (IDPs). A modified version of the UK Biobank Pipeline (Alfaro-Almagro et al. 2018) is used to extract these IDPs. Accordingly, this chapter also provides details of the processing pipelines for each modality as well as the necessary modifications made such as generalising it to accommodate data from all the scanners.

Inter-scanner variability for IDPs is mapped in a number of ways including comparing it to within-scanner, within-subject variability and between-subject

variability across different imaging modalities. This allows interesting questions to be answered such as "How robust are the IDPs extracted from one modality compared to another?" or "To what extent does acquiring data from the same vendor mitigate between-scanner variability, if at all?". The matter of how this variability can be reduced will be the subject of the subsequent chapter.

4.1 Introduction

A challenge with quantitative imaging is the lack of reproducibility of repeated measurements. As we saw before, a key part of the problem stems from the fact that images reflect macroscopic views of tissue and measurements frequently reflect indirectly the microscopic quantities of interest. A number of image processing and modelling steps are needed to map the measurements to features of interest. Here we explore in a comprehensive manner how inter-site variability is propagated through these processing steps.

The field of neuroimage processing and analysis has had explosive growth over the years. A number of methods underlying specific software packages are very commonly used in neuroimage processing for extracting features. Examples include FreeSurfer (Desikan et al. 2006, Fischl et al. 2004) (regional surface area, volume thickness), FSL (Smith et al. 2004) (multi-modal image processing), SPM (Friston 2007) (multi-modal image processing), AFNI (multi-modal image processing), fMRIPrep (Esteban, Markiewicz, Blair, Moodie, Isik, Erramuzpe, Kent, Goncalves, DuPre, Snyder et al. 2019) (brain region activation maps) and many more. More recently, a number of multi-modal pipelines have been developed for optimally combining one or more of these building blocks. These include the HCP pipeline (Glasser et al. 2013), the UK Biobank pipeline (Alfaro-Almagro et al. 2018) or more recent integrative environments such as QuNex (Ji et al. 2022). In combination, these enable the extraction of

Imaging-Derived Phenotypes (IDPs) which are summary measures and have the potential to be used as personalised brain signatures, e.g. for diagnosis, stratification or prognosis.

Despite the potential, inter-site differences such as scanning protocols, hardware, and magnetic field strength can contribute to lack of consistent quantifiability in these IDPs. Scans of the same individual obtained from different scanners can differ in somewhat unexpected and unpredictable ways, as variability propagates through the processing in nonlinear manners. Unwanted bias and variance are then introduced into the measurements and data become strongly associated with the acquisition site/scanner rather than with true biological variability.

Examples have been presented before. For instance in (Takao et al. 2011), it was shown that the variability caused by scanning in two different scanners gave the wrong impression of actual brain volume change over a two year period. An increase in mean volume change was observed but this was found to be a result of scanner hardware and software upgrade rather than actual brain volume change. In (Reig et al. 2009), it was shown on average that the variability of volumetric brain data induced by a multi-scanner set up was as high as 17%. An additional example was demonstrated in (Fortin et al. 2018). Linear discriminant analysis (LDA), used to find a linear combination of features which separate the data into two or more classes, was performed on cortical thickness measures and showed that data points clustered almost perfectly by site illustrating the extent to which scanners induce bias. Examples of measures from diffusion MRI include the study performed by (Vollmar et al. 2010). Here it was shown that inter-site variability in FA can be between 5 and 15% in white grey matter areas yet differences of interest due to pathology, e.g. in diseases such as schizophrenia, are often of the order of 5% (Mirzaalian et al. 2016). Furthermore, the study in (Hainline et al. 2018)

demonstrated scanner-induced bias and variance in generalised FA measures in selected ROIs. It was shown that not every ROI is affected by the same bias and variance showing that the bias and variance induced by multi-scanner acquisitions has implications on global brain measures as well as regions which are more local.

In previous studies, a single or a few imaging-derived features were considered to assess variability across scanners from one or two modalities (Mirzaalian et al. 2016, Zhu et al. 2011, Han et al. 2006). Furthermore, a typical issue in mapping this variability is the lack of some reference to compare against. For example, in the study performed in (Mirzaalian et al. 2016), a method is proposed to harmonise the raw signal from diffusion data. It is assumed that two separate groups of individuals, scanned in different scanners, but who are matched for age, gender, handedness and socio-economic status should have similar diffusion profiles and these were therefore used as references for each other. Another example is in (Fortin et al. 2017) where a method is proposed to harmonise derived FA and MD maps from diffusion data. Similarly, the consistency of summary measures across participants matched across studies for age, gender, ethnicity, and handedness is used to asses the quality of the proposed harmonisation method. In these examples it is evident that a reliable and consistent reference is unavailable and the choice of matched subjects is a compromise.

In this chapter we solve these challenges in more comprehensive ways compared to what has been tried before. First, we are using a rich travelling-heads dataset that includes all vendors and much more scanners than previous studies. Second, because we have acquired within-scanner repeats, we can use within-scanner variability as a reference against between-scanner variability. Third, we map variability for thousands of multi-modal IDPs, extracted using a modified version of the UK Biobank pipeline (Alfaro-Almagro et al.

2018). We compare not only against within-scanner variability, but also against between-subject variability (as a proxy to biological variability). Finally, even if it is well known that acquiring data on different scanners induces bias and variability, what is less explored is whether or not the ordering of subjects for specific features remains consistent and to what extent. We perform analyses towards this direction as well and show that certain modalities are more robust in preserving cross-subject ranking than others.

4.2 Theory

For each of the acquired scan sessions, we used a modified version of the UK Biobank pipeline (Miller et al. 2016, Alfaro-Almagro et al. 2018) to extract thousands of multi-modal imaging features for which we subsequently map their variability. In this section, we present the main principles of the pipeline, which is used as a backbone for the data processing. We overview the modifications we performed in Methods.

4.2.1 UK Biobank Processing Pipeline and IDPs

Coherent and consistent data processing flows are needed to analyse all acquired data and achieve from distortion and motion correction, alignment for within-subject modalities to multi-modal feature extraction. The UK Biobank pipeline (Alfaro-Almagro et al. 2018) was specifically designed with that purpose in mind in order to process multi-modal data acquired by the brain imaging component of the UK Biobank (Miller et al. 2016).

Given the differences across vendors, the pipeline (originally designed originally for Siemens Skyra input data) had to be adjusted in various ways, as outlined in Methods, to allow processing of all available data. This resulted in a pipeline that was generally flexible, with the ability to accommodate input data from all the types of scanners we used which is scheduled to be released

publicly. Additionally, the pipeline was also augmented to account for new ways of processing. Yet, the main principles remain the same and we overview them here.

The pipeline takes input data from each of the modalities, performs preprocessing, such as distortion correction, brain extraction, template normalisation, and generates IDPs. The IDPs are collections of multimodal features, including volumes of tissue types, cortical and subcortical volumes, measures of microstructure in white matter and structural connectivity, iron deposition proxies in grey matter, and functional connectivity properties. An overview of the features extracted from each modality is shown in Figure 4.1. In total the pipeline is capable of generating over 4350 features. A summary of the processing steps for each of the acquired modalities is now provided.

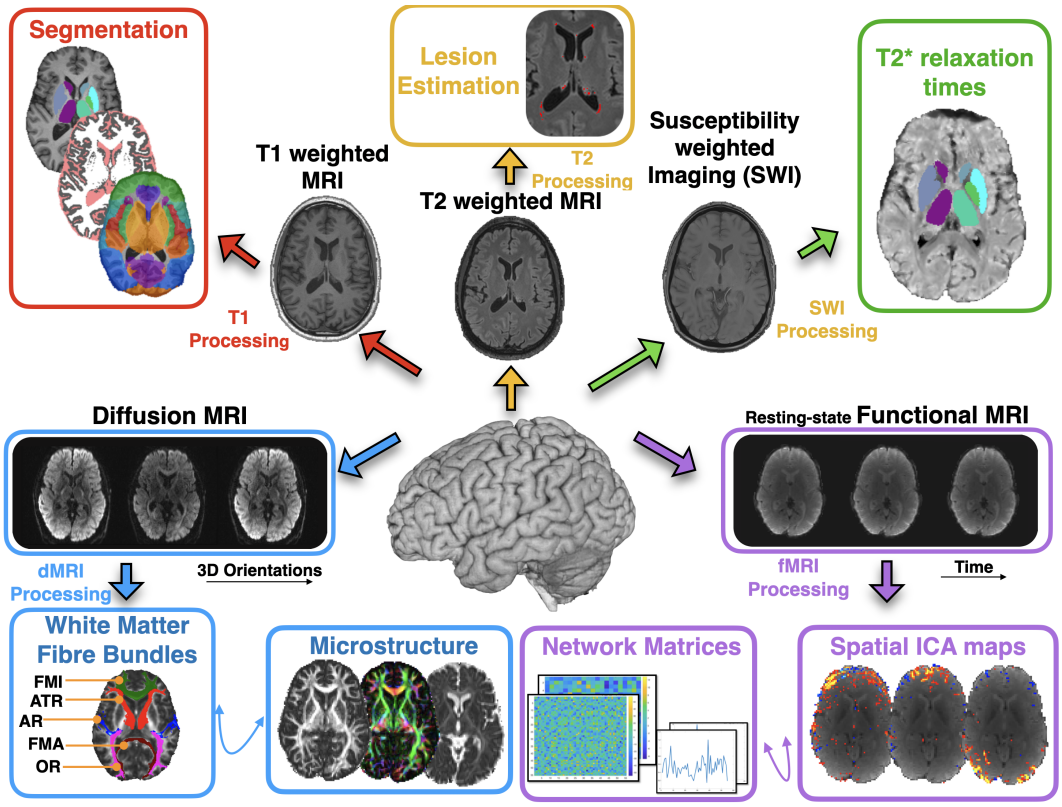

Figure 4.1: *Overview of the features extracted from each modality. The data from each modality were processed using a modified version of the UK Biobank pipeline to obtain a comprehensive set of imaging features across all scanning sessions.*

4.2.2 T1-weighted Processing

The T1 processing pipeline begins by reducing the field of view of the original image. Specifically, it removes empty space around the head and also removes some of the neck. This ensures a more accurate and standardised initialisation of the brain extraction which follows immediately afterwards. It also improves the robustness and accuracy of subsequent registrations. After this, the warp which transforms the image from native T1 space to a standard reference space is estimated using FMRIB's Nonlinear Image Registration Tool (FNIRT) (Andersson et al. 2007) . The reference space is the $1mm$ resolution version of MNI 152 template. This warp is then inverted so that it takes data from standard space into native T1 space and is applied to a brain mask. The brain extraction is performed by applying this mask onto the original brain image.

In addition, a series of steps are performed to transform the actual T1-weighted original image into standard space. This is done using a compound transformation which is a combination of 3 separate transformations: 1) The transformation from distorted to undistorted space, corrected for gradient nonlinearities (when gradient nonlinearity information is provided by the scanner manufacturer), 2) the transformation from the full field of view image to the reduced field of view image and 3) the transformation from the reduced field of view image to standard space.

Following template normalisation and brain extraction, there are 3 different groups of features generated from the T1-weighted processing. The first group are derived from segmented global tissue measures such as total brain volume, total volume of WM, GM and CSF and several others. These are generated using the SIENAX tool (Smith et al. 2002, 2004) for cross-sectional measurements. The second group are cortical parcel volumes which are derived from quantifying the amount of tissue labelled as grey matter in 139 distinct regions

of interest. The ROIs are defined by combination of parcellations from several atlases: Harvard- Oxford cortical and subcortical atlases, and Diedrichsen cerebellar atlas. The grey matter is segmented using FMRIB's Automated Segmentation Tool (FAST) (Zhang et al. 2001). The third group are subcortical volumes of 15 structures segmented using FMRIB's Integrated Registration and Segmentation Tool (FIRST) (Patenaude et al. 2011). The T1-weighted processing is summarised in Figure 4.2.

4.2.3 T2-weighted Processing

For the T2-weighted processing, the T2-weighted-FLAIR is first linearly transformed into the space of the T1-weighted image. Then to extract the brain, the standard space brain mask (previously transformed into T1 space) is applied to the T2-weighted image. A segmentation of white matter hyperintensities is subsequently performed by Brain Intensity AbNormality Classification Algorithm (BIANCA) (Griffanti, Zamboni, Khan, Li, Bonifacio, Sundaresan, Schulz, Kuker, Battaglini, Rothwell et al. 2016) and then thresholded at a value of 0.8 to generate a single feature of lesion volumes. The classification of voxels as either legions or non-lesions is based on a k nearest neighbour algorithm with training data from the UK Biobank. Although the processing for these IDPs was performed, they were not used in the results due to the age of the cohort being low (mean 34 $\pm$ 9.4 years) and the subjects being healthy. This is in stark contrast to the mean age of the 2 datasets used in the original study: 75 $\pm$ 7 years and 67.4 $\pm$ 14.3 years, thus lesions are not expected in our participants and the lesion estimates are likely to be dominated by noise. The T2-weighted processing is summarised in Figure 4.3.

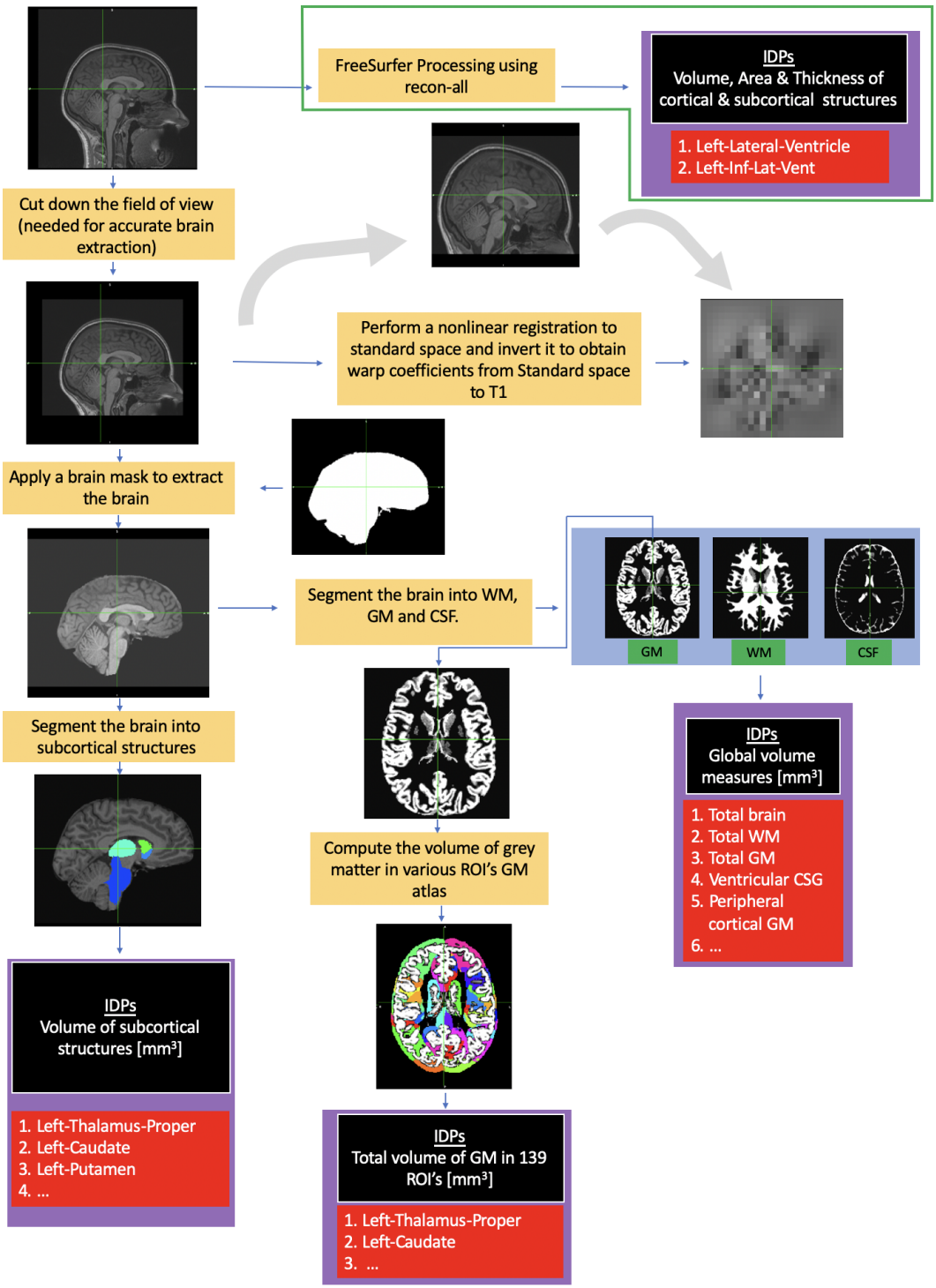

Figure 4.2: *Flowchart of T1-weighted processing showing the steps taken from input of a raw image to quantitative summary measures. Steps added to the original pipeline are outlined in green.*

4.2.4 SWI Processing

The SWI processing begins by combining data from individual coils into one image using a root-sum-of-squares approach, in case individual coil data are provided (e.g. for Siemens scanners). When data are already combined (e.g.

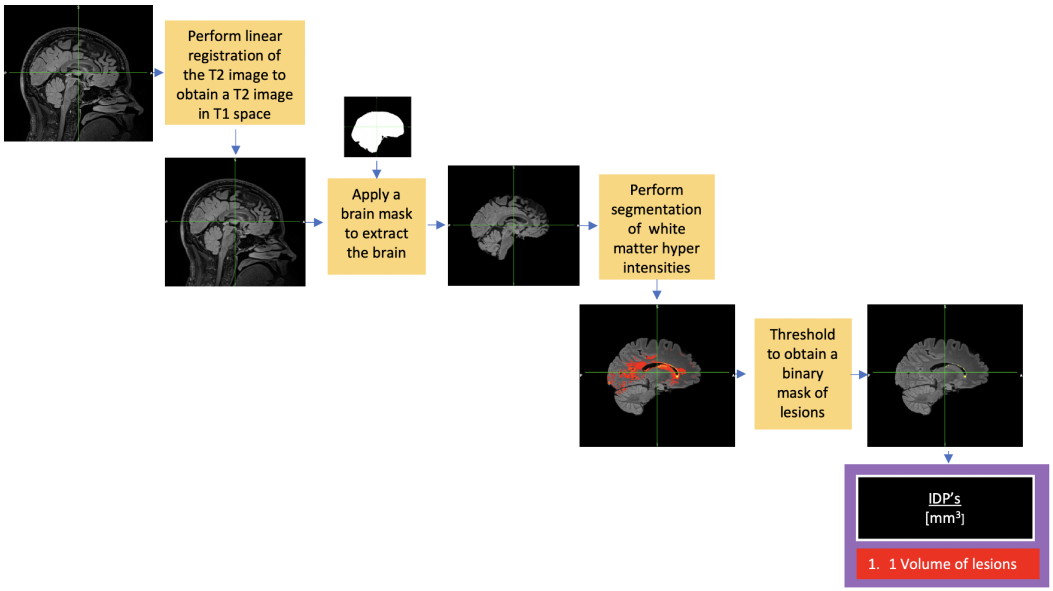

Figure 4.3: *Flowchart of T2-weighted processing showing the steps taken from input of a raw image to quantitative summary measures.*

typical for non-Siemens scanners) this step is not performed. Once combined, R2* is computed using the log of the ratio of the two echo time images scaled by the echo time difference and T2* is calculated as the inverse of R2*. Then a linear transformation is performed to take the data from SWI space and bring it into T1 space. A brain mask, transformed from the T1 space to the SWI space is then applied to the T2* image to obtain only the T2* values within the brain. The generated IDP's are the T2* values in the 15 sub-cortical structures segmented by FIRST during the T1-weighted processing. A venogram generation series of steps is also followed. The phase images are first high pass filtered before a complex image is generated by summing the complex data from each coil. This gives a single filtered phase image which is multiplied by the total magnitude image to get a venogram image with enhanced appearance of veins. The phase and magnitude images used for this process are from the second echo since it has greater venous contrast. The SWI processing is summarised in Figure 4.4.

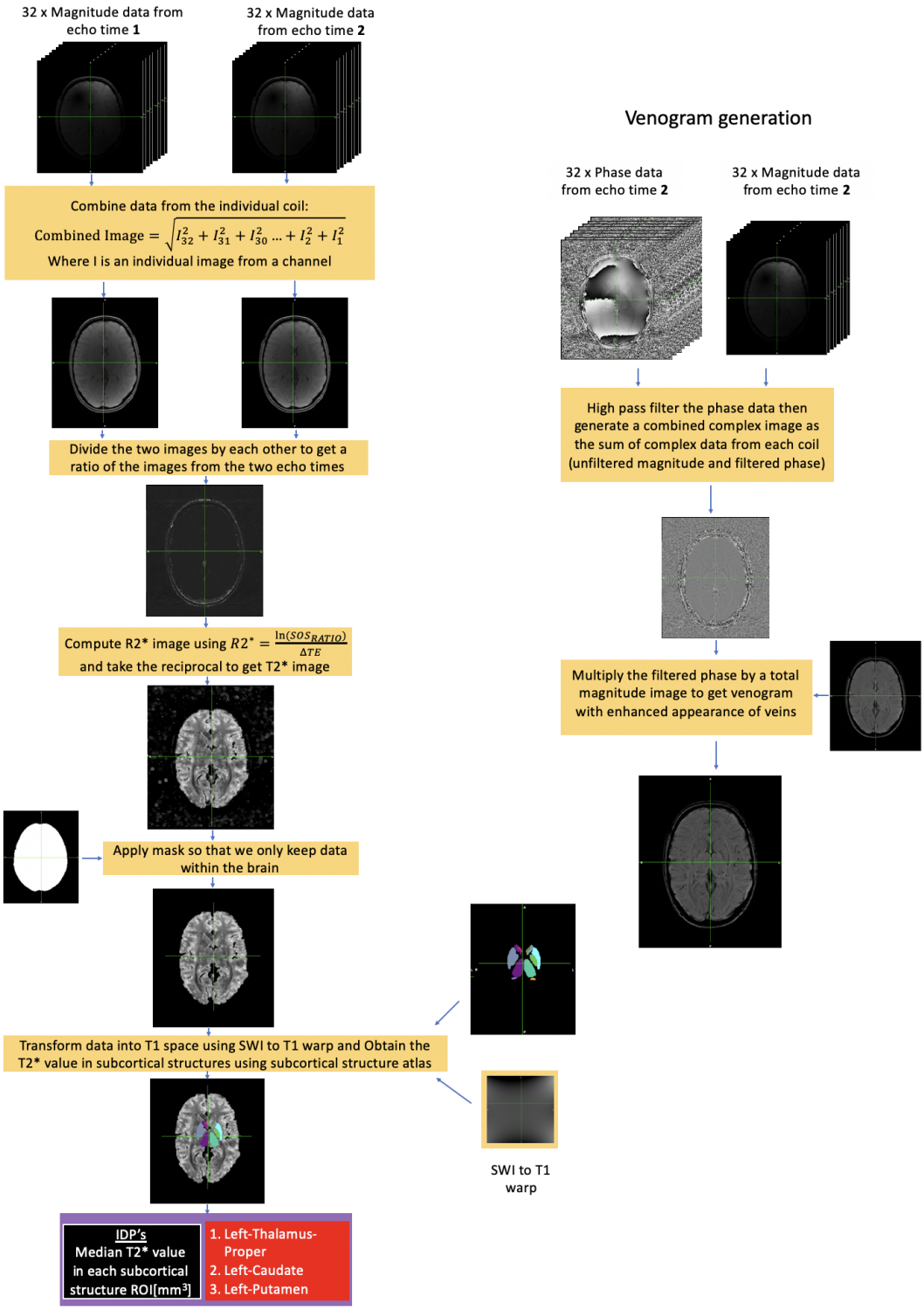

Figure 4.4: *Flowchart of SWI processing showing the steps taken from input of a raw image to quantitative summary measures.*

4.2.5 dMRI Processing

The first step of the dMRI processing pipeline is to correct for distortions and subject motion. This is done using the eddy tool (Andersson & Sotiropoulos 2016, Andersson et al. 2016). This accounts for eddy currents, subject motion

and also corrects susceptibility induced distortions using an off-resonance field which was calculated earlier (see section 3.3.5 in the previous chapter). After this, the DTIFIT tool (Basser et al. 1994) is used to apply a diffusion tensor model to give standard measures such as FA and MD. The data is also fed into AMICO (Daducci et al. 2015), an approximate nonlinear solver for fitting the NODDI model (Zhang et al. 2012). This is a multi compartment biophysical model to obtain measures on different types of microstructural properties, such as neurite density and fibre orientation dispersion, which are meant to be more specific than the simpler DTI metrics.

These microstructural features (DTI and NODDI) are summarised within white matter ROIs and their mean value in each ROI is extracted. ROIs are defined in two ways: a) Using a white matter atlas, specifically a set of 48 standard space tract masks defined by the JHU Template (Mori et al. 2005, Wakana et al. 2007). To reduce contamination from partial volume, these ROIs are further filtered using an FA-skeleton (obtained through the TBSS pipeline (Smith et al. 2006)), so that only the "core" of each tract is depicted in the values. B) Using subject-specific tractography.

Before the tractography can be performed, the data first needs to be fed into a modelling framework to estimate fibre orientations, using the BEDPOSTX (Behrens et al. 2007, Jbabdi et al. 2012, Hernández et al. 2013) tool. Using the BEDPOSTX output, probabilistic tractography is then performed by probtrackx (Behrens et al. 2007, Hernandez-Fernandez et al. 2019) using protocols from the Autoptx tool (De Groot et al. 2013). This generates 27 tracts that define subject-specific ROIs. Mean values of microstructural features within these tractography-based ROIs are also returned as features. The dMRI processing is summarised in Figure 4.5.

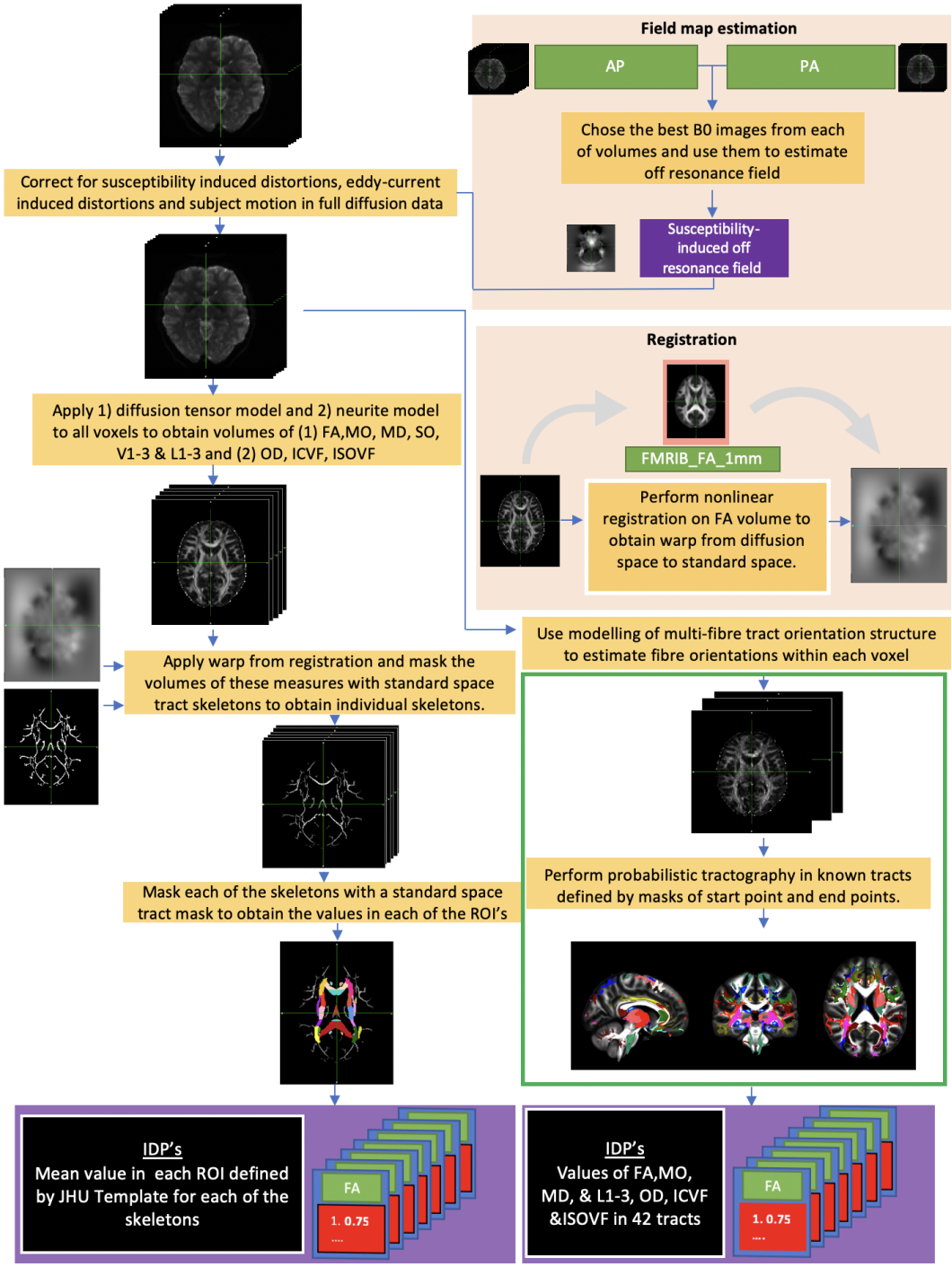

Figure 4.5: *Flowchart of dMRI processing showing the steps taken from input of a raw image to quantitative summary measures. Steps added/modified to the original pipeline are outlined in green.*

4.2.6 rfMRI Processing

A reference volume is first generated from the existing time series data. This is an image chosen from one of the first 5 images which correlates the most closely with the others. This is used for subsequent alignment to other modalities and motion correction. The field map generated from the spin-echo images is used here to correct for EPI distortions in the reference image. These corrections are done by the FEAT pipeline in FSL (Smith et al. 2004).

Independent Component Analysis (ICA) is then used to decompose the data into resting-state networks, components consisting of a spatial map and a time series. The components corresponding to noise are removed using FMRIB's ICA-based X-noiseifier (FIX) (Beckmann & Smith 2004, Griffanti et al. 2014, Salimi-Khorshidi et al. 2014). The training data that has been used for FIX consists of 40 UK Biobank training data sets.

Once FIX denoising has been performed, functional networks are mapped. Group level spatial maps generated from 4100 subjects from the the UK Biobank are mapped into the subject space, using the cleaned data. These define gray matter nodes that are used for connectivity estimation. The generated IDPs consist of connectivity (i.e. correlations of time series between pairs of nodes) and the node amplitudes of each regressed ICA component. The total number of rfMRI IDPs (3432) is the sum of the upper diagonals of two 21×21 matrices (partial and full correlations of non-artefactual components from a 25-dimensional group ICA - 21 independent components kept), the upper diagonals of two 55×55 matrices (partial and full correlations of non-artefactual components from a 100-dimensional group ICA - 55 independent components kept) and $21 + 55$ node amplitudes of the surviving components. The rfMRI processing is summarised in Figure 4.6.

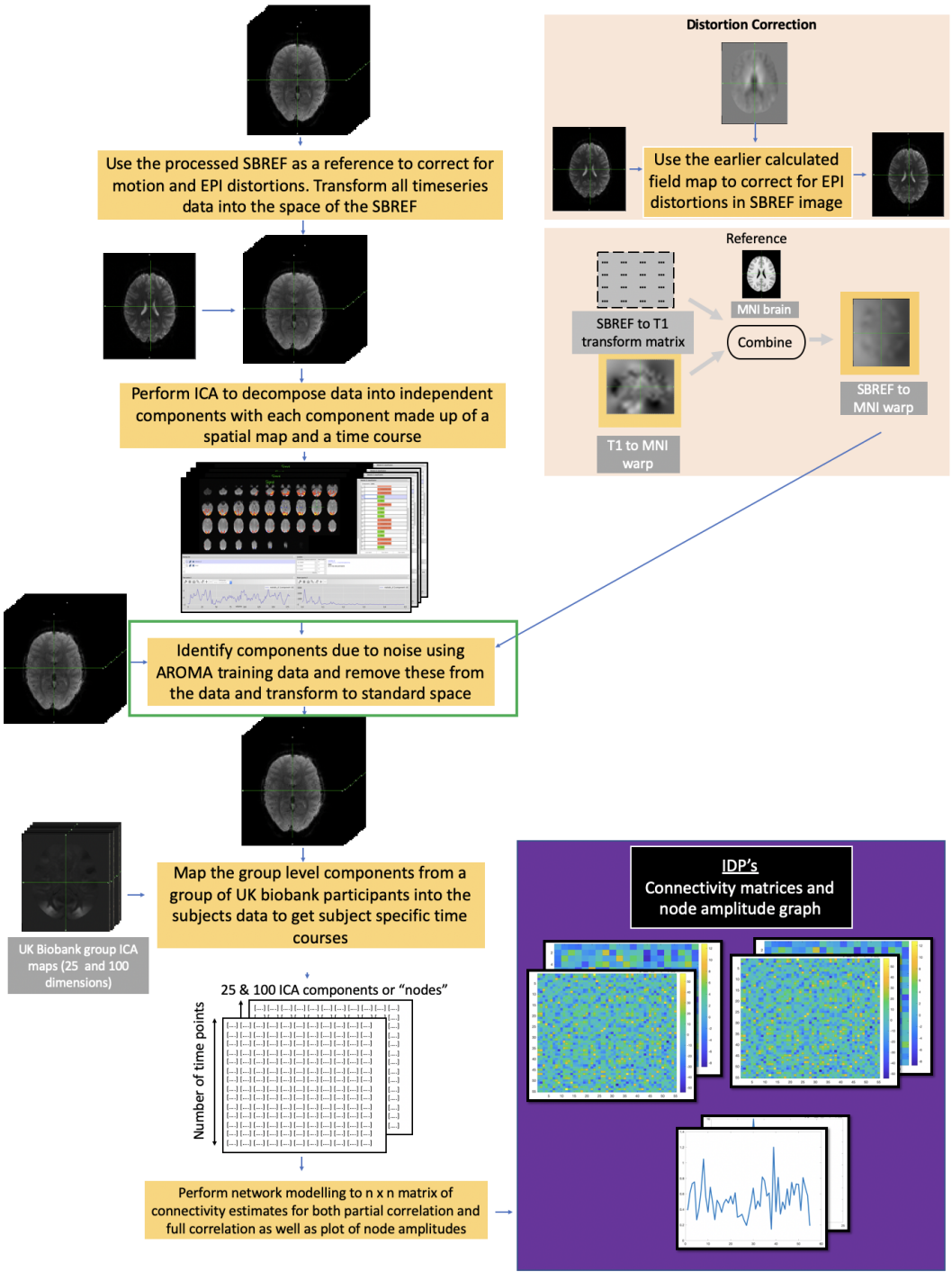

Figure 4.6: *Flowchart of rfMRI processing showing the steps taken from input of a raw image to quantitative summary measures. Steps added/modified to the original pipeline are outlined in green.*

4.3 Methods

4.3.1 Alterations Made to the UK Biobank Pipeline

The original pipeline, initially specific to the UK Biobank datasets and scanner vendor was modified in a number of ways to handle scans from the other vendors or augmented to add additional functionality and derive a greater number of IDPs. Components of the pipeline which have been thus modified appear with a green border in Figures 4.2 - 4.6

Data onboarding

A restructuring of the original UK Biobank pipeline was needed so that the pipeline could accept as inputs data from non-Siemens scanners. Apart from NIFTI data, accompanied json files are used and the pipeline reads acquisition meta-data from these json files. Given however that dicom-to-nifti conversion does not guarantee similar fields of json files to be populated for the same features from different vendors, we had to accommodate changes accordingly. At the time of the study we had to manually calculate the effective echo spacing and total read out time as outlined in section 3.3.5, however this is now performed automatically by dcm2niix (Li et al. 2016). These modifications facilitate bringing the data to a BIDS format (Gorgolewski et al. 2016), which will enable wider sharing of the data in the near future, in repositories such as https://openneuro.org For conversion to NIFTI, the dcm2niix (Li et al. 2016) tool v1.0.20181114 GCC7.3.0 was used.

Handling corrections for gradient nonlinearities

Gradient nonlinearities cause geometric distortions and they are larger at further distances from the iso-center of the scanner. Even if vendors provide gradient nonlinearity corrections for both 3D and 2D acquisitions on the scanner, the UK Biobank pipeline allows this correction to be performed as a post-acquisition step, due to inconsistencies identified between the 3D and 2D

corrections in the Siemens scanners (Alfaro-Almagro et al. 2018). In order to be able to perform these corrections "offline", a vendor proprietary file is needed that characterises the spatial distribution of these inhomogeneities. We have no access to these files for the non-Siemens scanners and therefore the online corrections provided by the vendors were performed for the Philips and GE data. To allow this in the pipeline , we had to add the option of toggling off the gradient distortion correction (GDC), which occurs at various preprocessing steps and gets merged with all warps. Specifically, toggling on and off GDC influences how warp fields are estimated between modalities and from a subject to standard space, and necessary changes were implemented to accommodate such change.

Handling SWI (magnitude and phase data from different vendors)

The pipeline was also generalised to accommodate the different SWI data from the scanners. For the Philips and GE scanners, the coil data had already been combined so we made the step which combines coils optional. Furthermore, for the GE data, the complex phase image had already been combined and filtered so we additionally made these steps optional.

Computing spin-echo fieldmaps

As mentioned in section 3.3.5, the effective echo spacing for the dMRI images and the rfMRI images also had to be calculated for each scanner at the time of the study (This is no longer needed as these measures are these are now automatic outputs of dcm2niix (Li et al. 2016)) , as this was used to correct for EPI distortions using spin-echo fieldmaps. These would then be used to derive the readout times. The readout times are necessary for correct application of the off-resonance field maps during corrections for susceptibility induced distortions.

Toggling pipeline generated SBREF and acquired

For rfMRI data, a single band reference volume was acquired (SBRef) on the Siemens scanners. This has higher between tissue contrast and the same geometry as the rest of the time series data. This is used for subsequent alignment to other modalities and motion correction. This is not available for all scanners and also a less aggressive Multiband acceleration was used for non-Siemens scanners that preserved more contrast. Therefore, we added an option to toggle this on or off based on whether it was acquired on the scanner or not. In the cases where it was not acquired, a reference volume was generated by the pipeline from the existing time series data. This is an image chosen from one of the first 5 images which correlates the most with the others.

Adding additional features to the Pipeline

We also augmented particular parts of the pipeline to take advantage of new developments. Specifically: 1) In the T1-weighted processing, we added FreeSurfer reconstructions to allow for cortical thickness, area and volumes using the subject derived parcellations rather than standard space atlases. 2) In the dMRI processing we performed two changes. We replaced the legacy Autoptx (De Groot et al. 2013) protocols with the more state-of-the-art XTRACT (Warrington et al. 2020) tractography, which has 42 tracts and has been more widely evaluated. We also replaced AMICO (Daducci et al. 2015) for NODDI model fitting with cuDIMOT(CUDA diffusion modelling toolbox) (Hernandez-Fernandez et al. 2019). NODDI nonlinear fitting is very computationally expensive and slow and cuDIMOT accelerates it by using GPUs to do the optimisation; while AMICO does a series of linear approximations to the non-linear model, which reduce accuracy, while being slower than cuDIMOT. 3) In the rfMRI processing, we added the option of unsupervised denoising using ICA-AROMA (Pruim, Mennes, van Rooij, Llera, Buitelaar & Beckmann 2015). This is an alternative to FIX which has been trained with Siemens UK Biobank data and may not be optimal for non-Siemens scanners, compared to AROMA which is agnostic to vendor.

4.3.2 Assessing IDP Cross-session Similarity

Once the modified UK Biobank pipeline was ran across all 80 scanning sessions in our database, we first looked into cross-session similarity of IDP patterns for each of the four individual subjects that had within-scanner repeats. A procedure similar to that described in Section 3.3.7 was used. For each subject, and for each scan session, a vector containing a certain set of IDPs was constructed. These were defined a-priori and consisted of all IDPs excluding those derived from rfMRI and T2. These were omitted because of these modalities are significantly more noisy compared to the ones included and their inclusion resulted in trends being indiscernible. This IDP vector, representative of the session, was correlated with the respective IDP vectors of all other sessions (11 in total, including between-scanner and within-scanner sessions) of that subject. IQMs were z-scored across sessions and the IDP vectors were magnitude-normalised, before calculating the pairwise correlation of sessions. This was performed to explore the relationship between between-scanner and within-scanner similarity of IDPs in the same subject.

4.3.3 Testing for Scanner Effects Across IDPs

Scanner effects were statistically tested using repeated-measures ANOVAs. For each IDP, the null hypotheses was tested that there was no difference in the group means of IDP values from data acquired in repeated measurements across different scanners (i.e. repeated measures were the between-scanner repeats). Each group consisted of 10 values, representing 10 measurements of the same IDP obtained in each of the 10 subjects. As there were multiple IDPs and therefore tests performed, multiple comparison correction was done and adjusted the threshold for rejecting the null hypothesis using false discovery

rate (FDR=5%, within each test). The adjusted threshold was determined according the procedure proposed by Benjamini and Hochberg (Benjamini & Hochberg 1995). The estimation for scanner effects was performed 3 times: a) using the data from all 6 scanners (i.e. 6 repeated measures) b) using the data from the Philips scanners only (i.e. 2 repeated measures) c) using the data from Siemens scanners only (i.e. 3 repeated measures).

4.3.4 Mapping IDP Between-scanner Variability

We first used the within-scanner data, from subjects that had repeated scans, to determine the within-scanner coefficient of variation (standard deviation of measurements divided by mean) of each IDP. This allowed us to gauge which group of IDPs were most robust in the absence of site effects. We then mapped between-scanner variability and compared it to a number of references, as explained below.

With respect to within-scanner variability

We used the within-scanner repeats as a baseline for assessing between-scanner variability. The assumption is that within-scanner variability reflects a minimum scan-rescan variation that can be observed for a subject, when scanner hardware, software and operator stay the same, and is therefore mostly driven by thermal noise. We therefore calculated a bias measure (equation 4.1) and a relative variability measure (equation 4.2). The bias measure quantified the difference between the average (median) across between-scanner measurements and the average across within-scanner measurements as a percentage of the latter. The relative variability measure reflected the difference between the variability (interquartile range) of between-scanner measurements and within-scanner repeats as a percentage of the latter.

$$\text{Bias } = \frac{\text{Median }_{\text{between-scanner}} - \text{ Median }_{\text{within-scanner}}}{\text{Median }_{\text{within-scanner}}} \times 100 \tag{4.1}$$

$$\text{Relative Variability} = \frac{\text{IQR}_{\text{between-scanner}} - \text{IQR}_{\text{within-scanner}}}{\text{IQR}_{\text{within-scanner}}} \times 100 \tag{4.2}$$

With respect to between-subject variability

We also compared the between-scanner variability of IDPs from a single subject to measures of between-subject variability for the same IDPs. In an ideal scenario, it is expected that measurements from the same subject across different scanners vary considerably less than measurements from different subjects, which reflect biological variability. To explore this for the set of IDPs considered here, we used two between-subject variability metrics: a) One that reflected IDPs from scans of the 10 subjects in our cohort acquired from the same scanner, b) A close-to-population-level biological variability metric, by considering IDPs from 1000 subjects randomly-chosen in the UK Biobank and scanned in the same scanner (Siemens Skyra). For the within-cohort metric, we used the variability across the 10 subjects when scanned on a Siemens Prisma scanner. For the UK Biobank derived metric, we accessed data using UK Biobank Project 43822 (PI: Sotiropoulos, Univ. of Nottingham).

4.3.5 Assessing the Preservation of Subject Ranking between-scanners

In addition to bias or variance increase in metrics due to inter-scanner effects, another potential detrimental effect for quantifiability in studies is the inability to preserve subject ranking across scanners. Using our data we explored

this effect. For each IDP, we explored the consistency in subject ranking as depicted between-scanners.

To do so, for each of the 6 scanners and for each IDP, the 10 subjects were ranked, leading to 6 ranking vectors of length 10 per IDP. The first vector simply had the numbers 1-10 in ascending order and the order of the 10 subjects in the other 5 vectors was determined with respect to their order in this first vector. The consistency in subject ranking for each IDP was calculated by taking the average Spearman correlation of all the possible pairwise combinations of the 6 vectors. We did this for 3 cases: a) considering only the Siemens scanners b) considering only the Philips scanners d) considering all the scanners.

To get a lower bound of the consistency of rankings, we computed a null representing correlations that would have been obtained from random rankings of the subjects between the scanners. This was done by simulating 6 random vectors each containing integers 1-10 and calculating the average Spearman correlation across all the possible pairs. This was repeated 1000 times to give a distribution of correlations. We defined the null region as the values between the -75^{th} and the 75^{th} percentile of that distribution.

4.4 Results

4.4.1 Cross-session IDP Similarity

We computed correlation matrices depicting similarities of IDPs across different scanning sessions. The contrast between the two distinct regions shown in Figure 4.7 confirms the expectation that within-scanner repeats have a higher correlation with each other than the between-scanner measurements. This il-

lustrates a greater consistency in values of IDPs derived from within-scanner measurements compared to those derived from between-scanner data. Differences appear to be less prominent for subject 14482, as the IDP values derived from the within-scanner measurements in the particular scanner (Achieva) were less consistent than those from the other scanners. A possible reason for this could be the the large amount of motion associated with the scans from the Achieva. Results from Figure 3.7 in the previous chapter showed that the within-scan repeats acquired in the Achieva had the largest amount of subject motion which would explain this.

A noticeably high correlation exists with the 3 Siemens scanners in subject 14482. It is worth noting that these scans were acquired on the same day within a 5-hour window which is by far the shortest timeframe among all the scans in the Figure. In contrast, the Siemens data for subject 13192 was acquired over a 3-day period, subject 14229 was over a 3 month period and subject 14230 was over a 2 day period.

Nevertheless, on average we can observe that that IDPs from within-scanner repeats are more similar to each other; and that the two Siemens Prisma sessions are more similar to each other than any other scanner session, followed by similarity to the Siemens Trio sessions. Interestingly, the Philips Achieva sessions seem to be closer to the GE MR750 and Siemens Trio (both narrow-bore scanner), rather than the Philips Ingenia, which is a wide-bore scanner.

4.4.2 ANOVA Results

T1-weighted and SWI IDPs

We tested for mean scanner effects in repeated-measures ANOVAs. Figure 4.8A shows p-values for different groups of IDPs based on structural modal-

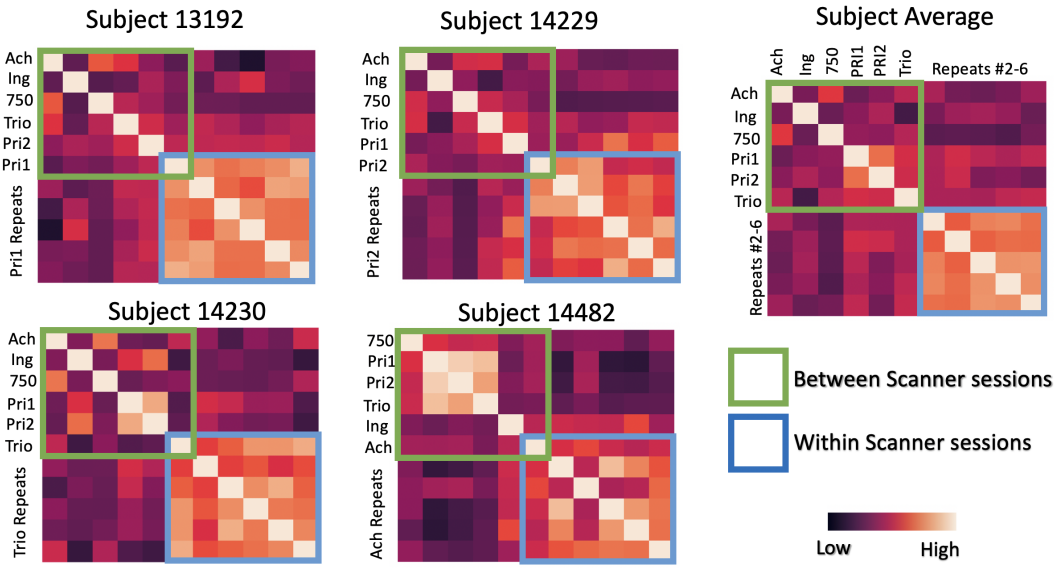

Figure 4.7: *Correlation matrix depicting correlations of IDPs across sessions. The data used is from the 4 subjects for whom repeat scans were acquired. The range of the colour bar has been set from the* -90^{th} *the* 90^{th} *percentile of values.*

ities. There is smaller evidence for differences driven by site effects for data acquired from the same vendor. For a number of these IDPs this remained true even when data from a combination of vendors was considered. The cortical parcel volumes and cortical thicknesses in particular showed a greater number of differences in group means for within-vendor data as well as between-vendor indicating that these measures were less consistent across different scanners.

dMRI IDPs

Figure 4.8B shows the distribution of p-values for repeated measures ANOVA performed on diffusion parameters estimated regionally for white matter tracts. Here we use the results from ROI-based IDPs only to represent what would typically be expected from diffusion quantitative measures. There is limited evidence for differences in the group means of values when data is compared across only Siemens scanners. For Philips scanners, we see a greater number of p-values representing a significant difference in group means but we see yet a greater number of significant p-values when the comparison is done across the different vendors and more site effects than the structural IDPs.

rfMRI IDPs

IDPs representing functional connectivity strength between different brain regions did not show a high level of reliability across sites (see following Figures). Therefore, we show results for node amplitudes across scanners in Figure 4.8C. A small number of p-values showed significant differences in group means when all scanners were considered, but in general, these IDPs did not show variations in the mean of the groups acquired from different scanners.

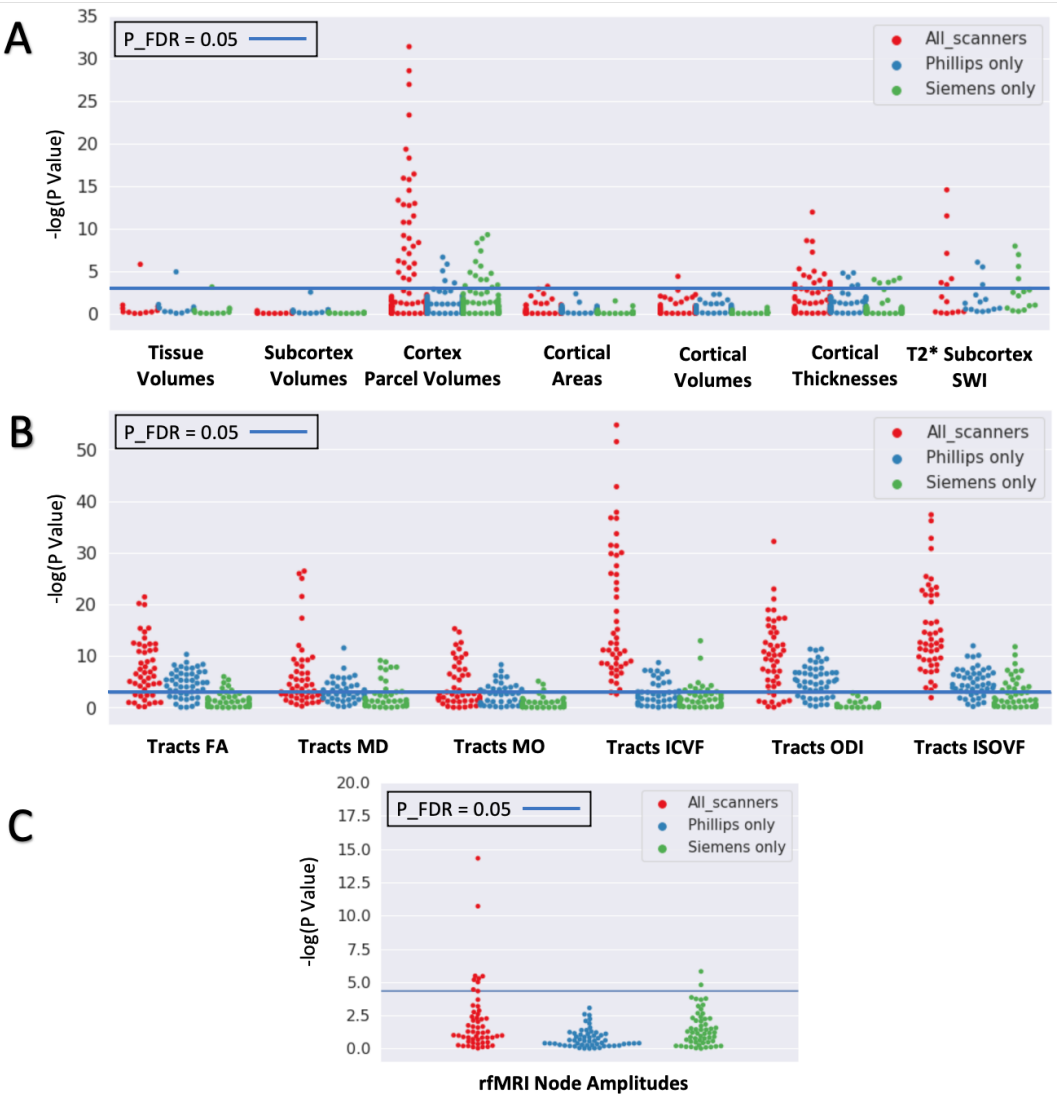

Figure 4.8: *A distribution of log-transformed p-values from repeated measures ANOVA for A) Structural data B) Diffusion data and C) Functional Data. The solid horizontal line represents the threshold for significance and its value is the p-value equivalent to FDR=5%. Cases have been considered where IDPs have been extracted from only Siemens scanners, only Philips scanners or all scanners.*

4.4.3 Mapping IDP Variability

First, we mapped within-scanner and between-scanner variability for all IDPs. The variability of IDPs within the same scanner was assessed using the scan-rescan data of each subject. Figure 4.9 shows the coefficient of variation (CoV) (standard deviation of measurements divided by mean) of IDPs taken across 6 measurements acquired A) within the same scanner B) acquired on each of the 6 different scanners for a representative subject chosen because their IDP values matched most closely with the others subjects.

The 3432 rfMRI IDPs include connections between all edges (connection between nodes) including those which represent weaker or false positive connections. We therefore considered in these plots only the strongest 5% edges from the 100-dimensional group ICA using full correlation as a measure of functional connectivity. The strongest edges were determined by ranking the edges according to mean connectivity strength across within-scanner repeats and retaining the same top 5% across all comparisons. Nevertheless, the rfMRI connectivities had much higher coefficients of variation than the rest of the IDPs thus they appear outside the range of visualisation.

The magnified inserts of Figure 4.9 indicate that most of the imaging features have higher between-scanner CoV than within-scanner CoV . Qualitatively, the structural IDPs obtained from FreeSurfer processing show the smallest difference in variability of within- vs between-scanner repeats and seem to be more robust to site effects.

It is worth pointing out that we identified a consistent bias in some diffusion IDPs obtained from the Ingenia sessions. FA and MD values were lower and higher respectively for a number of regions in the Ingenia scans compared to the other scanners for the same subject. This trend was particularly evident in some of the subjects; in fact, the magnitude of the trend seemed to cor-

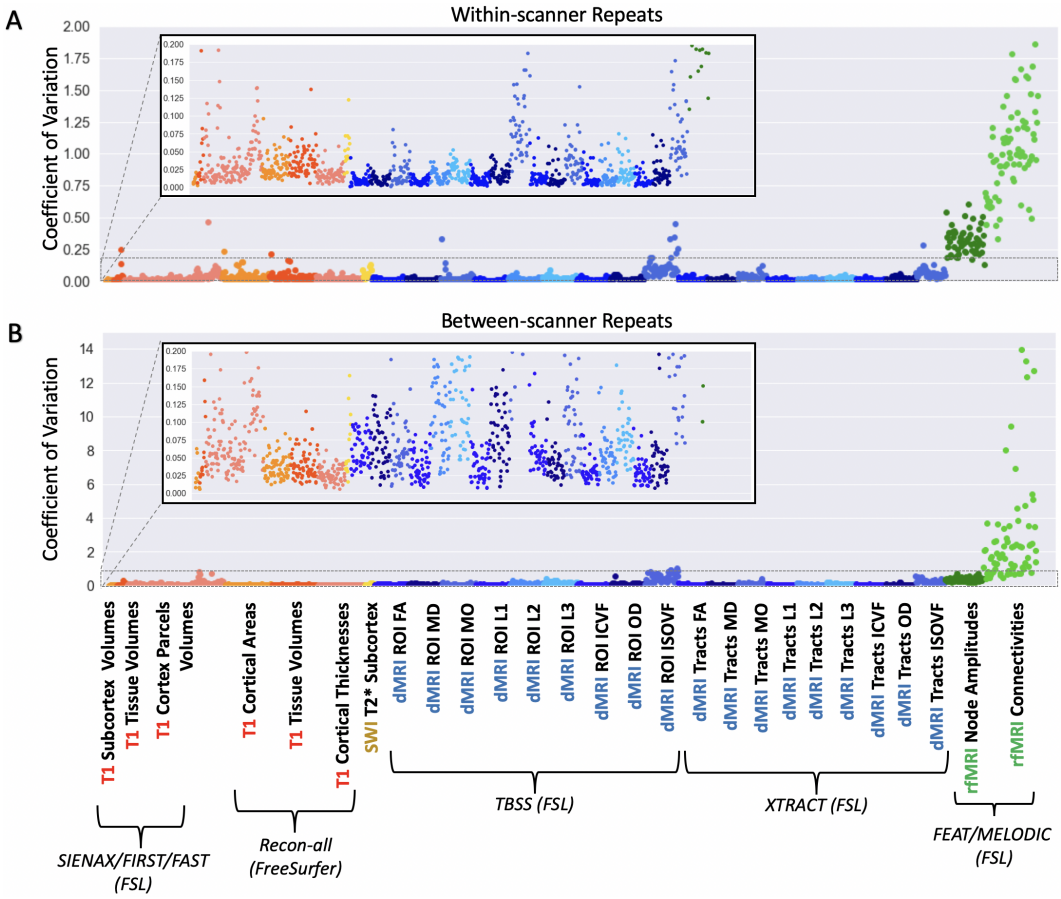

Figure 4.9: *The CoV of IDPs taken across 6 measurements acquired A) within the same scanner(acquired on the Siemens Prisma FMRIB scanner) B) acquired on each of the 6 different scanners for a single subject . The IDPs have been colour-coded with respect to their modality and they have also been grouped according to how they were processed. *The GE data has been excluded from ICVF, OD and ISOVF IDPs, as due to the single-shell protocol, NODDI IDPs cannot be obtained from the GE data.*

relate with head size (the larger the head the higher this Ingenia bias). In the Appendix 4.6 we show this issue and we discuss it in greater detail. We tried a few protocol variations in the Ingenia dMRI scans (e.g. changing the multiband factor and in-plane acceleration combinations, details in Appendix). However, the issue did not fully resolve when a multiband factor ≥ 2 was used, pointing to challenges in simultaneous multi-slice dMRI on the wide-bore Ingenia. For these reasons, we decided to not include the Ingenia dMRI IDPs in the following figures of this chapter, which therefore represent the 5 more consistently-behaved scanners for dMRI (a version of Figure 4.9 with Ingenia dMRI IDPs excluded is shown in the Appendix, Figure 4.16).

Figure 4.10 shows examples of between-scanner variability for specific IDPs for a single subject. We look at the values obtained from each scanner for the following features: the volume of the left hippocampus, the total volume of CSF in the brain, the mean FA in right temporal part of the cingulum tract and the amplitudes from a chosen connectivity node. From this narrow selection of IDPs, apart from the node amplitude, we see that values derived from scanners of the same vendor are more consistent with each other whereas those obtained from different vendors frequently appear to be more distant. For example, in the the boxplots for the total volume of CSF and the left hippocampus volume, it is the values obtained from the 3 Siemens scanners which lie closest to the median. Conversely, this trend is less prominent for the node amplitudes where it is seen that ordering of the scanners appears more shuffled.

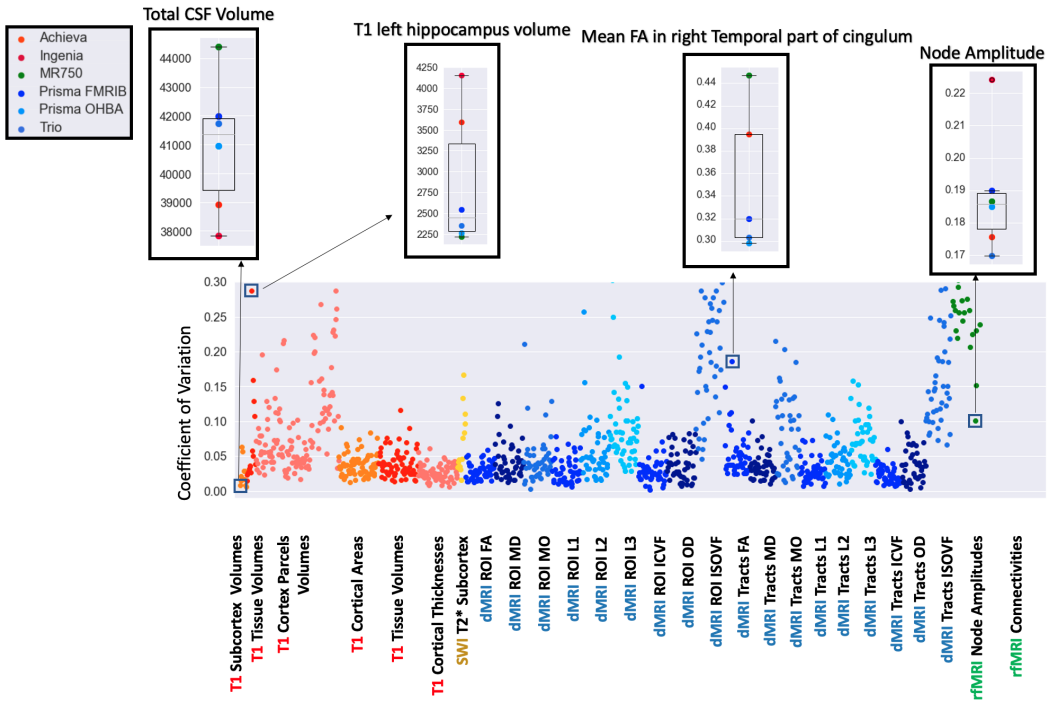

Figure 4.10: *Explicitly showing the variation obtained from different scanners for a selection of IDPs. Considered IDPs are the volume of the left hippocampus, the total volume of CSF in the brain, the mean FA in right temporal part of the cingulum tract and the amplitude from a chosen connectivity node.*

So far, we have qualitatively mapped differences in IDP variability of within

vs between-scanner repeats. We subsequently **explored the IDP between-scanner variability** more quantitatively against different references.

With respect to within-scanner variability

Figure 4.11 shows a plot of the relative variability for a single subject and for an average across 4 subjects. We have shown this for 3 different cases: a) Comparing values from Siemens scanners alone b) comparing values from Siemens and Philips scanners c) comparing values from all scanners. To aid visualisation, we include only ROI-based IDPs for diffusion.

Note, from this point forward, the results from tract-based ROIs have been omitted from these figures. To aid visualisation, we only show ROI-based IDPs. Furthermore, both sets of IDPs pass a similar message when compared to IDPs from other modalities. We will reintroduce tract-based IDPs in the next chapter when we directly compare ROI-based IDPs and ROI-based IDPs.

In all cases, we see that the vast majority of values are positive (for the 4 subject average and for all scanners, only 11.3% of considered IDPs had negative values), indicating that imaging features had much higher between-scanner than within-scanner variability, reaching on average up to 10 times more. The between- vs within-scanner variability for each IDP category was 71% more (median values for the average across all 4 subjects) for structural IDPs, 78% more for SWI, 153% more for dMRI and 36% more for rfMRI. We also observe that as we begin to compare across different vendors, the relative variability increases, complementing what was observed in Figure 4.8.

Table 4.1 shows the relative variability of between- vs within-scanner repeats for finer IDP categories in detail. We see that there are distinct changes in relative variability from one imaging modality to another. Features derived from structural MRI show the least relative variability. This is followed by SWI and then dMRI derived features. Some interesting patterns can be observed in the

Table 4.1: The median relative variability difference of between- vs within-scanner repeats for IDP categories (average of 4 subjects).

IDP Category	Median Relative Variability
T1 Subcortex Volumes	305%
T1 Tissue Volumes	74.5%
T1 Cortical Areas	70.0%
T1 Cortical Volumes	53.6%
T1 Cortical Thicknesses	71.8%
SWI T2* Subcortex	78.0%
dMRI ROI FA	215%
dMRI ROI MD	166%
dMRI ROI MO	89.7%
dMRI ROI L1	126%
dMRI ROI L2	219%
dMRI ROI L3	215%
dMRI ROI ICVF	204%
dMRI ROI OD	181%
dMRI ROI ISOVF	296%
dMRI Tracts FA	399%
dMRI Tracts MD	156%
dMRI Tracts MO	93.8%
dMRI Tracts L1	110%
dMRI Tracts L2	242%
dMRI Tracts L3	268%
dMRI Tracts ICVF	118%
dMRI Tracts OD	102%
dMRI Tracts ISOVF	93.0%
rfMRI Node Amplitudes	48.6%
rfMRI Connectivities	24.5%

dMRI IDPs that agree with intuition. For instance, FA values are more prone to site effects than MD values, probably reflecting the fact that FA is a higher-order statistic (variance) compared to the MD (mean). Another example is that the variability increases with smaller diffusivities, i.e. L1 has less relative variability than L2 and less than L3, as the smaller the diffusivity value is, the more prone to noise. The rfMRI IDPs show comparable consistency even to those from T1-weighted imaging, however, as noted before, we have only considered the top 5% of edges and also small differences in variability point to high within-scanner variability for these features. The rfMRI IDPs also exhibit a large bias, demonstrated in the next Figure.

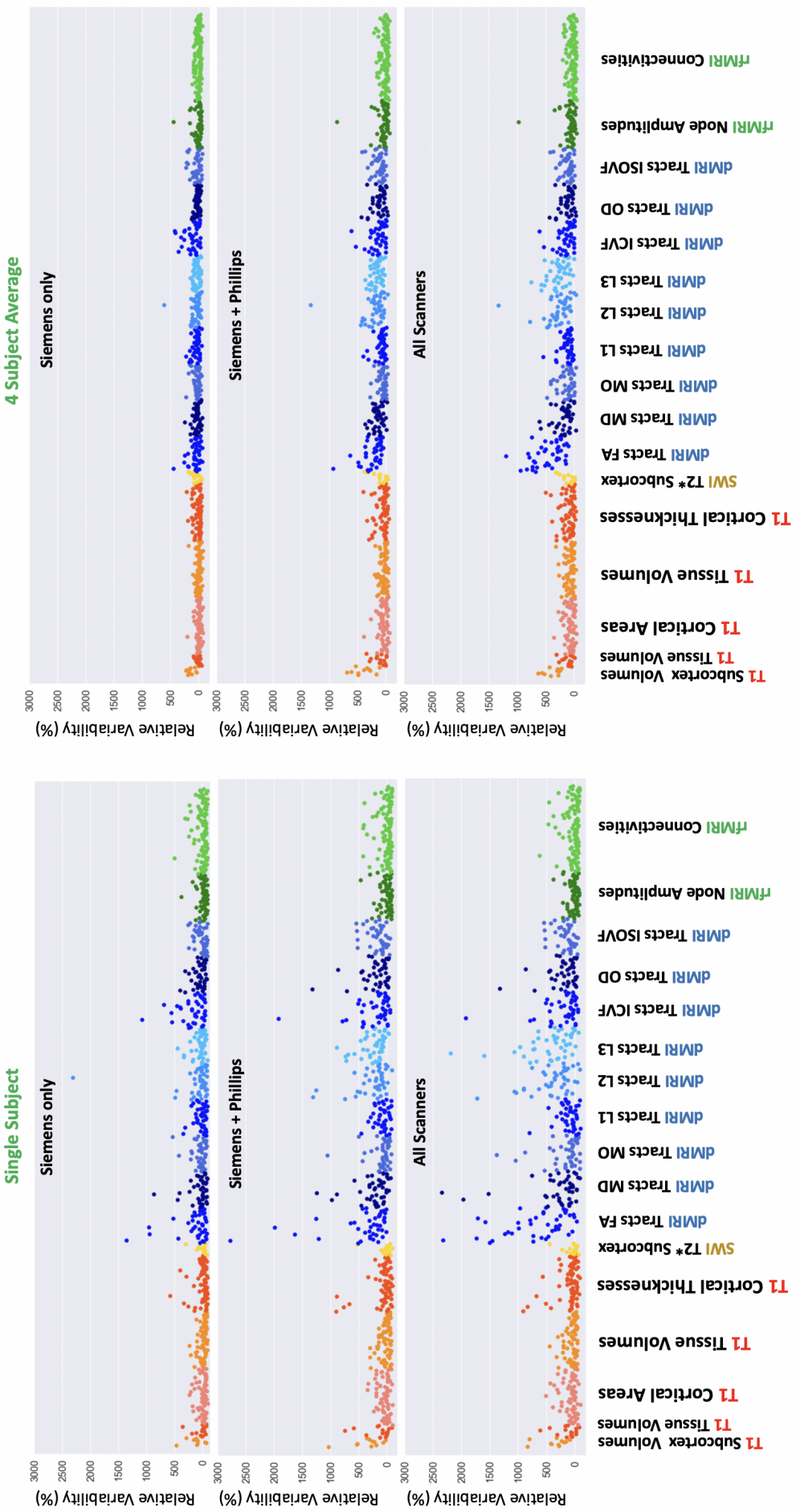

Figure 4.11: *The relative variability of imaging-derived measures comparing the interquartile range of between-scanner measurements for different scanners and within-scanner repeats of the same subject, reflecting a single subject (left column) and the average across 4 subjects (right column). First, second and third row reflect the 3 Siemens, 5 Siemens and Philips and all 6 scanners for the between-scanner repeats, respectively.*

Figure 4.12 shows a comparison of the median values for each IDPs when considering between-scanner and within-scanner repeats, providing a relative measure of bias. We can see that the trends oscillate around zero (which corresponds to no bias), but for a given subject and most IDPs, the median of between-scanner measurements is different to the median of within-scanner repeats by up to 15%. When averaging, these biases values drop, they can still however be in the order +/- 5%. The rfMRI IDPs are an exception, and they do show large differences in the average values measured across scanners compared to scan-rescan measurements.

In summary, the different trends observed for different IDP groups show that these differences also depend on the imaging modality and the processing used to extract the features, and both can have a considerable effect on the magnitude of this variability.

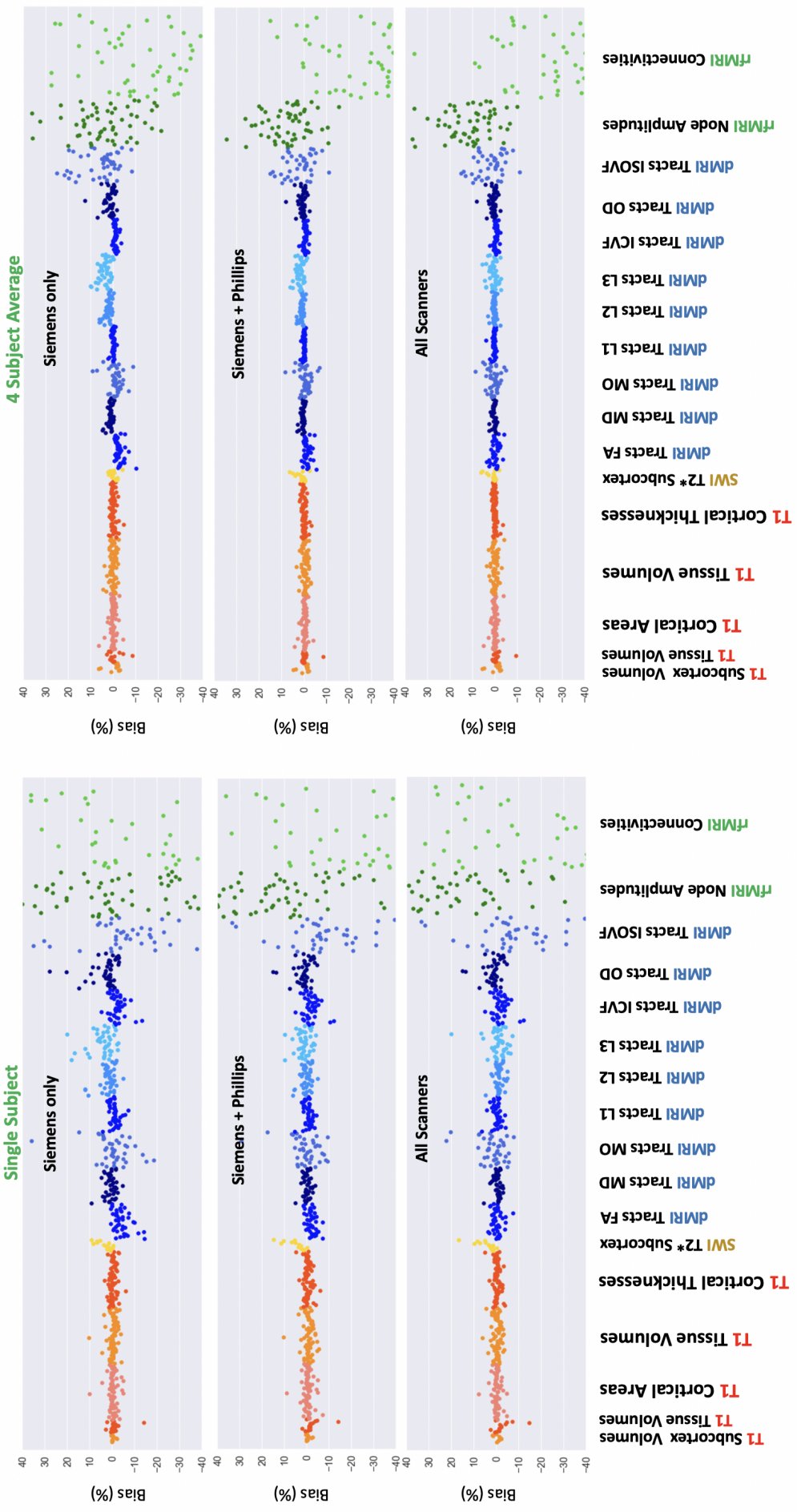

Figure 4.12: *The bias of imaging-derived measures comparing the median range of between-scanner measurements from 6 different scanners and 6 within-scanner measurements, reflecting a single subject (left column) and the average across 4 subjects (right column). First, second and third row reflect the 3 Siemens, 5 Siemens and Philips and all 6 scanners for the between-scanner repeats, respectively.*

With respect to between-subject variability

An interesting comparison is how variable are features extracted from scans of the same person in different scanners compared to the same features extracted from scans of different persons. We therefore compared the variability of IDP values obtained from the same subject being scanned in the 6 different scanners against references that represent between-subject variability. We defined these in two ways: a) IDP variability across the 10 subjects from our cohort being scanned in the same scanner. The scanner chosen for this was the Siemens Prisma FMRIB scanner due to the fact it most closely resembled the the scanner used in the UK Biobank brain imaging project (Siemens Skyra with 32 3T with a standard Siemens 32-channel RF receive head coil). We will call this "within-cohort variability", b) population-level IDP variability, using 1000 subjects from the UK Biobank, all scanned in a Siemens Skyra scanner. We will call this "UK Biobank variability". Figure 4.13 shows comparisons across these pools of variance for different IDP categories, where also the within-scanner variability is presented for reference. As expected, the boxplots corresponding to within-scanner repeats are the least variable, while the ones corresponding to 1000 different subjects are the most variable.

Results from statistical tests showed that for most cases there was a significant difference between the CoV's of IDPs obtained from within-scanner repeats and those from all other scenarios. Interestingly, there are a number of cases where the means of the CoV's of between-scanner measurements are not statistically different from those obtained from IDPs within the cohort. These are the T2* in subcortex regions derived from SWI acquisitions and the MD, L2, L3 and ISOVF from white matter tracts determined using regions of interests. As can be seen from the ISOVF, there are even cases where the CoV's of IDPs obtained from between-scanner measurements of the same subject are not statistically different from those obtained from 1000 UK Biobank subjects suggesting that for some IDPs the within-subject, between-scanner variability

is as large as the within-scanner, between-subject variability.

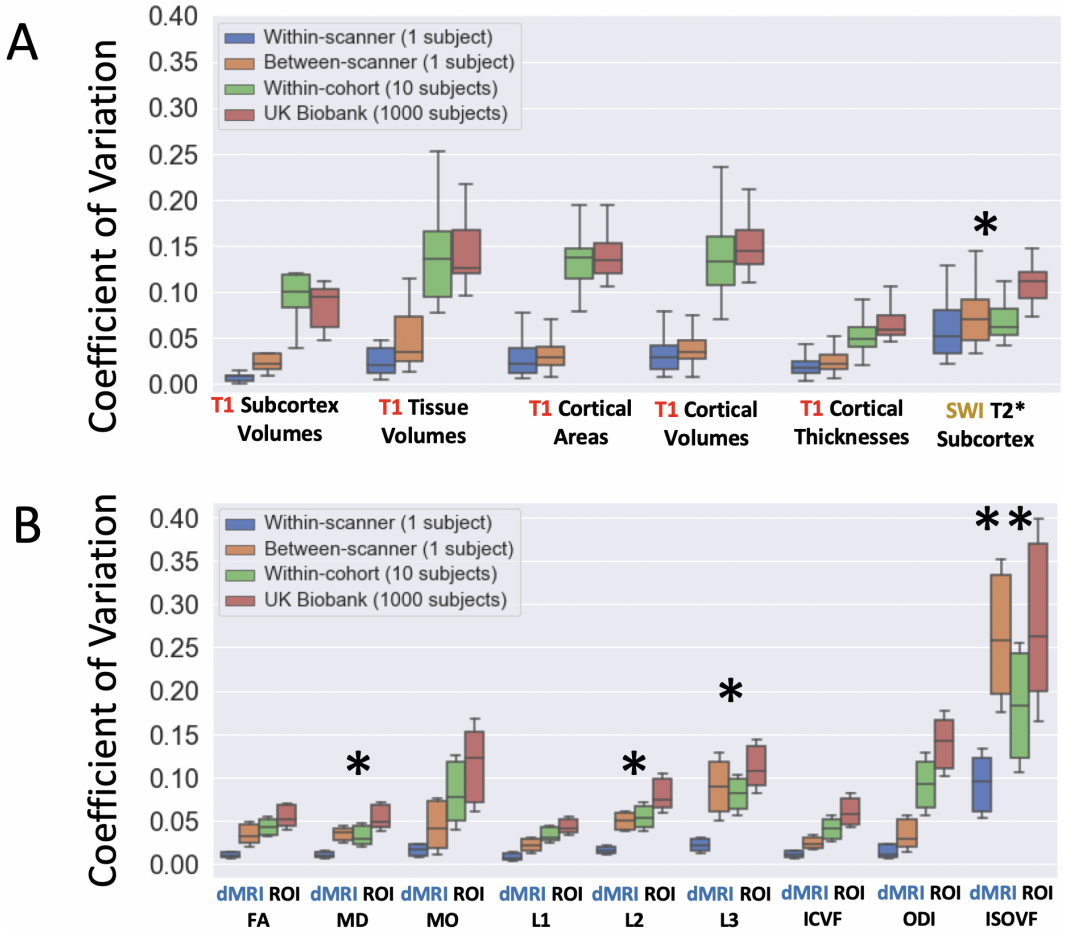

Figure 4.13: *Plots showing the distribution of coefficients of variation for different types of IDPs. 1* indicates the absence of statistical difference in between-scanner variability and within-cohort variability. 2 *'s indicate the absence of statistical difference in between-scanner variability and biological variability as determined by 1000 UK Biobank subjects. A non-parameteric Mann-Whitney U-test was used.*

4.4.4 Between-subject Ranking Consistency Across Scanners

For each IDP we explored the consistency in subject ranking as depicted across scanners. The results are shown in Figure 4.14. A value of 1 means perfect consistency, i.e. all subjects are ranked in the same way when using IDPs from different scanners. The red region is a "null consistency" regime that we identified using simulations of random rankings. We see that ranking is preserved

more for scanners from the same vendor but this ranking becomes more inconsistent when we include scanners from different vendors. Regardless, there are only a few categories of IDPs that are close to the ideal consistency described above. Furthermore, the extent to which ranking is preserved varies depending on the imaging modality. Between-subject ranking is preserved most for IDPs from structural imaging modalities, followed by diffusion, and least for functional modalities.

This is true even for IDPs which appeared to have been robust against site effects in Figure 4.13. A specific example is the cortical areas derived from T1-weighted measurements which show a non-significant difference between the CoV's derived from within-scanner-measurements and those derived from between-scanner measurements suggesting strong robustness to site effects. While this is true for the degree of variability, ranking can be affected and situations can be observed where a person A is measured to have a larger brain region than person B in one scanner, yet in another scanner the opposite is the case. This is seen in Figure 4.14 where we see the cortical areas derived from T1 measurements having a correlation value of less than 1.

Figure 4.14 shows specific examples of IDPs where the relative ranking between subjects presented. We see in Figure 4.14A that, despite global shifts in the values, for the total volume of grey matter, the ranking of the subjects is largely consistent across all scanners. This is in contrast to Figure 4.14B, where we can see that the the ranking values of the fractional anisotropy in the body of the Corpus Callosum are preserved to a lesser extent. This is consistent with the fact that dMRI as modality is more susceptible to scanner effects than T1-weighted imaging.

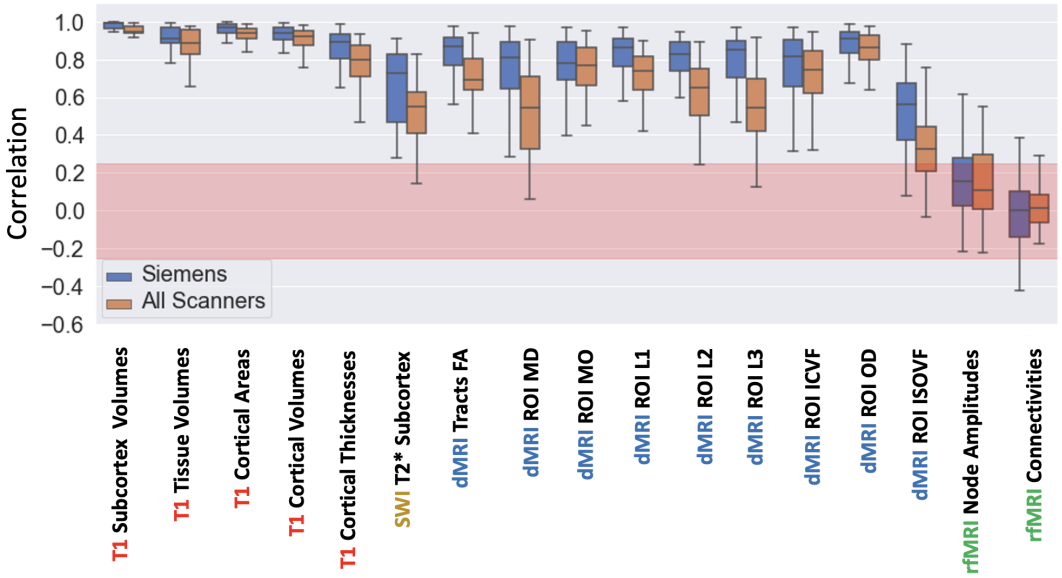

Figure 4.14: *Between-scanner consistency of subject ranking for different categories of IDPs. All ten subjects are ranked using the IDPs from each scanner and the relative ranking is correlated across scanners. The average of ranking correlations across IDP categories is depicted in this Figure. Trends have been shown for cases where only Siemens scanners have been considered and all scanners considered.*

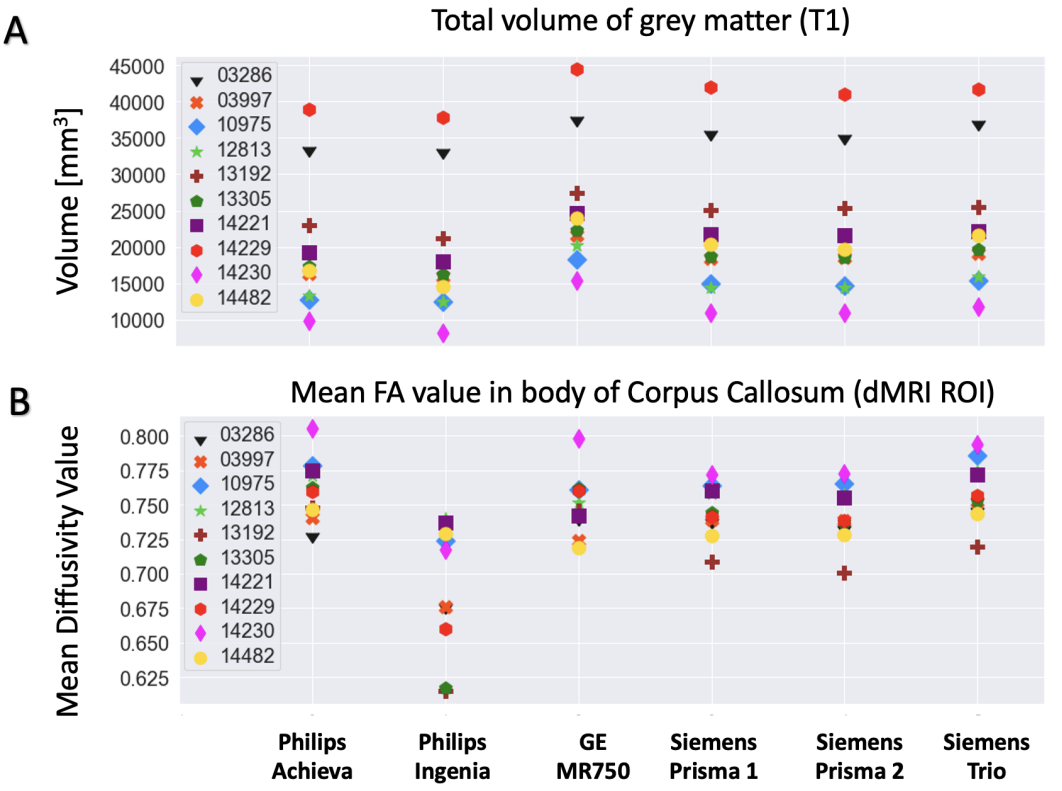

Figure 4.15: *Selected examples showing the extent to which between-subject ranking is preserved across scanners. We show here the results for (A) the total volume of grey matter which show ranking being largely preserved compared to (B) the Mean FA value in the body of the Corpus Callosum.*

4.5 Discussion

We have used a novel brain MRI harmonisation database to map between-scanner variability for a large range of multi-modal neuroimaging-derived features. We have found that when scanning a subject multiple times, between-scanner repeats induces variability which can be up to 5-10 times more the variability of within-scanner repeats, while bias can be in the order of 10-15%. Importantly, for a number of features this between-scanner variability can be of the same size as between-subject variability. We also found that consistency in subject ranking across scanners can be compromised relatively easily, particularly for certain modalities and features.

Our study maps the need for harmonisation in much a more comprehensive manner than before. By using a modified and augmented version of the UK Biobank pipeline, we extracted thousands of multi-modal IDPs and assessed their behaviour. The data we acquired of the same subject scanned on the same scanner allowed us to investigate broadly which modalities yielded the most consistent and reliable results and specifically the groups of IDPs within those modalities. We saw that IDPs derived from T1-weighted imaging are the most consistent. This was followed by IDPs derived from dMRI yet even within these IDPs there was a spectrum of variabilities depending on the type of measure. The IDPs derived from rfMRI were the least consistent. These trends are consistent with the extent to which these modalities are affected by noise and reflect the findings of other studies (Duff et al. 2021) comparing the variabilities of multi-modal data. We have also shown that the least between-scanner variability is observed when using scanners from the same vendor, as anticipated. Introducing different vendors increases the variability in IDPs and also decreases consistency in ranking of subjects across scanners.

Previous work has reported similar trends as the ones reported here. For in-

stance, structural imaging metrics were the most reproducible of the imaging features we present and this is consistent with past findings. High repeatability of these metrics has been shown across a range of segmentation approaches (de Boer et al. 2010), across multiple sites (Jovicich et al. 2006) and across scanners of varying magnetic field strength (Fujimoto et al. 2014). Cortical Areas and volumes derived from FreeSurfer have been to shown to even be robust to different acquisition sequences (Knussmann et al. 2022). It is worth noting that among the various groups of structural IDPs, a previous study (Duff et al. 2021) has shown that cortical area and thickness as derived from FreeSurfer are more robust than the grey matter volumes which were estimated for 139 ROIs and is in agreement with our findings.

For diffusion related metrics, previous studies have shown that generally, NODDI parameters have larger between-subject variation than DTI measures. The coefficient of variation for ISOVF has been observed to be consistently the largest among diffusion measures (Chung et al. 2016), which is in agreement with our results. Unlike structural measures, diffusion measures have been also found to be less robust against other factors such as magnetic field strength (Farrell et al. 2007) as DTI based contrasts particularly suffer from form poor accuracy as a result of low SNR. At lower field strengths such as 1.5T (Farrell et al. 2007) the discrepancy in reproducibility between of FA measures compared to MD measures is more stark than at 3T (Chung et al. 2016). The cited results still show FA measures to exhibit a higher degree of variability which is in agreement with our results.

For functional IDPs, it has been reported previously that test-retest reproducibility is a limiting factor (Castellanos et al. 2013) and also explains the small relative variability values we found. The results we have presented demonstrate that difference in variability of between- vs within-scanner repeats in rfMRI was low, as within-scanner variability was already high. Other

studies which performed similar analysis (Duff et al. 2021) pointed out that IDPs reflecting pairwise connectivity (as well as node amplitudes) do not show a high level of reliability across sites. Furthermore in the study performed by (Jovicich et al. 2016), significant inter-site differences in connectivity scores were found.

Compared to previous studies, in our work we have considered much more variability in scanners than before. We have also obtained within-scanner repeats so that we can compare directly with scan-rescan measurements. This provides a more reliable and consistent reference to be used during the assessment or development of harmonisation approaches. It mitigates the need for using methods such as subject matching which can be inherently challenging. The large number of travelling heads and variety of scanners used will also allow harmonisation approaches to be assessed based on their ability to retain consistency in between-subject ranking which, based on the literature, is relatively unexplored.

Our study had some limitations. For the dMRI IDPs, we identified a consistent bias in the Ingenia sessions and we opted to keep these excluded from the analyses. In doing so, we effectively excluded a worst-case scenario dataset and present a lower-bound for IDP variability, which is still very significant. Furthermore, we have relied on a certain set of IDPs, which is tied to a certain pipeline and software tools. It is clear that the IDP variability also reflects robustness of the processing pipeline steps, as well as variability in data quality. This is exemplified in the difference we found in variability between atlas-based cortical parcel volumes and Freesurfer-based ones, which are segmentation based. The latter were less variable than the former, which indirectly shows that segmentation-based methods are more robust to registration-based methods. Nevertheless, regardless of the pipeline choices, this is still the most comprehensive mapping of inter-scanner effects on imaging-derived features

across multiple modalities.

The findings presented here have mapped the extent of the problems the lack of harmonisation can induce and we hope that they can become a useful resource and reference for future studies. The next chapter will focus on using the the dataset we have acquired as a test bed for harmonisation. We will see if, how and to what extent, the issues highlighted in this chapter can be mitigated by some of the most established harmonisation approaches in the field.

4.6 Appendix: Bias in Ingenia dMRI Sessions

As mentioned previously in section 4.4.3, FA and MD values were lower and higher respectively for a number of regions in the Ingenia scans compared to the other scanners for the same subject. This is shown in Figure 4.16A which contrasts the effects on a subject with a large Ingenia bias compared to a subject with small Ingenia Bias. Figure 4.16B shows images of the corresponding MD maps for each case. We tried a few protocol variations in the Ingenia dMRI scans as shown in Table 4.2. The issue did not fully resolve when a multiband factor ≥ 2 was used. We also explored combinations of different parallel imaging factors (Sense) and slice acquisition orders. In each case, there was no noticeable improvement from the default protocol (MB=3, in-plane SENSE=2). This may point to challenges in simultaneous multi-slice dMRI on the wide-bore Ingenia. The issues was mostly mitigated without multiband, but in that case TR was extremely high and the total acquisition time was prohibitive (more than 25 minutes). Figure 4.17 is a version of Figure 4.9 with the Ingenia data excluded.

Table 4.2: The various Ingenia protocol options attempted in order to remove bias.

Multi-band Factor	Sense Factor	Slice Acquisition Order
3	1.5	Sequential
2	1.5	Sequential
2	1.5	Interleaved

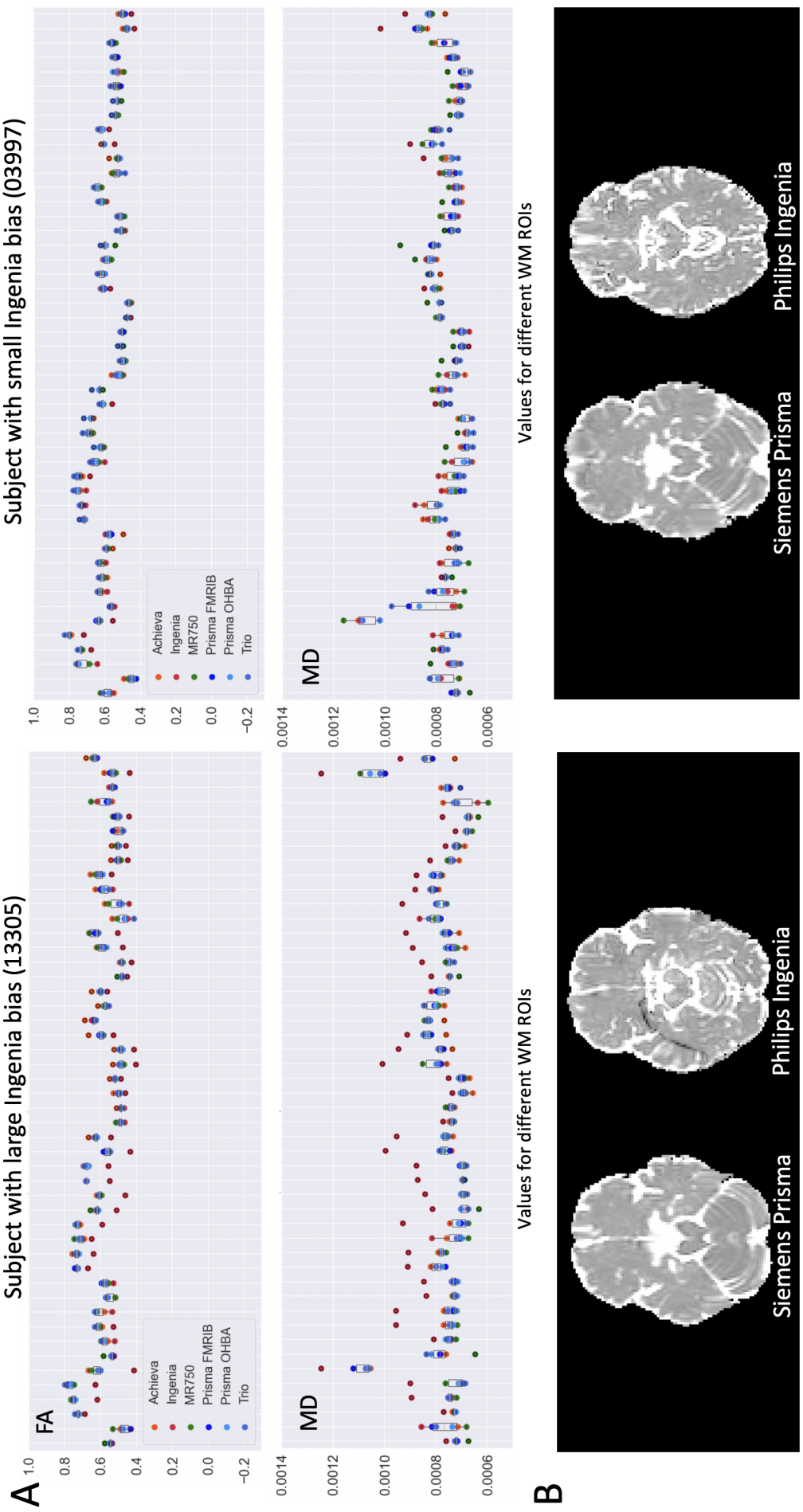

Figure 4.16: *Demonstration of issues for dMRI Ingenia data. A) Consistent biases were seen for some subjects (an example on the left), where the FA values were consistently lower for Ingenia scans compared to other scanners, and the MD values consistently higher (red dots). IDP values for a subject where this bias is less evident is shown on the right. B) Example of images of MD maps of subject with and without large Ingenia Bias and how this compares with an MD map of the same subjects in a different scanner.*

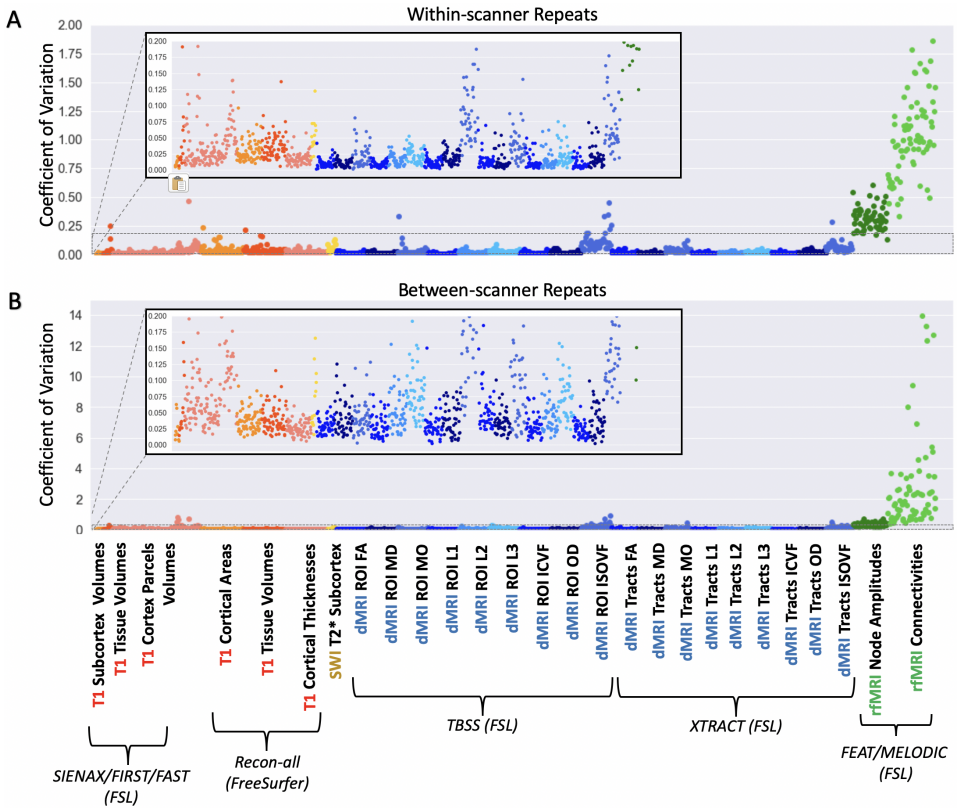

Figure 4.17: *The CoV of IDPs taken across 6 measurements acquired A) within the same scanner (acquired on the Siemens Prisma FMRIB scanner) B) acquired on each of the 6 different scanners scanners for a single subject excluding data from the Philips Ingenia. The IDPs have been colour-coded with respect to their modality and they have also been grouped according to how they were processed. *The GE data has been excluded from ICVF, OD and ISOVF IDPs, as due to the single-shell protocol, NODDI IDPs cannot be obtained from the GE data.*

Chapter 5

A Testbed for Evaluating Harmonisation Approaches

5.1 Introduction

In the previous chapter, we used our data resource to quantify the variability of imaging features of the same subject scanned in different scanners. Whilst the previous chapter highlighted the extent of the problem, this chapter explores different ways of addressing the problem of harmonisation.

One way to *harmonise* neuroimaging data is through approaches which *explicitly* aim to remove or mitigate scanner/site effects. These include frameworks such as ComBat (Fortin et al. 2017), Neuroharmony (Garcia-Dias et al. 2020) and the many others reviewed in Chapter 2. A challenge across these studies is the lack of objective or ground-truth references for evaluating the developed harmonisation approaches. For example, several prior works have used the approach of subject matching (Mirzaalian et al. 2016, Fortin et al. 2017) where it is assumed that two separate groups of individuals matched for characteristics such as age, gender, handedness and socio-economic, ethnicity, should have similar feature profiles and any difference between these factors is attributed to scanner-related inconsistencies.

A more direct approach for assessing the quality of harmonisation is to use within-scanner repeats. For example, in (Vollmar et al. 2010), 2 within-scanner repeats are used as a reference for assessing harmonisation reliability. The assumption is that a good harmonisation algorithm should reduce between-scanner variability of data or features to levels similar to within-scanner variability. Therefore, the smaller the difference between the coefficient of variation (or the greater the correlation) of between-scanner measurements and within-scanner measurements, the more effective the harmonisation approach is deemed to be. This approach has a more objective baseline as there is a reference of "gold standard" or "minimum variability", although the 2 within-scan repeats in (Vollmar et al. 2010) was a relatively low number of measurements to reliably establish this gold standard. The resource that we have developed is more ideally placed to do this as we have significantly more within-scan repeats, the variability of which can be assessed and used as a reference/target for post-harmonised between-scanner variability. We therefore propose our dataset as a testbed for evaluating such approaches.

It has been stated that a good preliminary step for achieving reproducible data is to first ensure harmonisation of scanning protocols across scanners (Chalavi et al. 2012). In a similar way, a prudent preliminary step before applying explicit harmonisation approaches is to have as much as possible the least variance and bias achievable. This can be primarily achieved through a thoughtful selection of processing tools and pipelines. It has been shown that using different tools to process the same data can produce significantly different results. For example, in (Botvinik-Nezer et al. 2020), the authors show that 70 different teams using different pipelines to analyse rfMRI data produce widely differing results on the same data. In (Schilling et al. 2021), when 42 independent teams were given diffusion data to process, the largest source of variability in the results was the processing tools which they used.

This variability exceeded the variability caused by differences in scanning protocol and the differences across the subject themselves. In (Griffanti, Rolinski, Szewczyk-Krolikowski, Menke, Filippini, Zamboni, Jenkinson, Hu & Mackay 2016), the results of using different artefact removal tools and the choice of the set of independent components indicated a lack of reproducibility between different analysis settings. It is therefore evident that there is a need to obtain optimal pipelines which yield the most reproducible and consistent results. We term this *implicit harmonisation*.

Figure 5.1 shows pictorially the main principles of implicit vs explicit harmonisation, with the anticipation that the latter is more effective in reducing between-scanner variability, but typically harder to do, as it requires bespoke algorithmic frameworks. While the former can be a more straightforward task of optimising preprocessing steps in a way that minimises propagation of data variability.

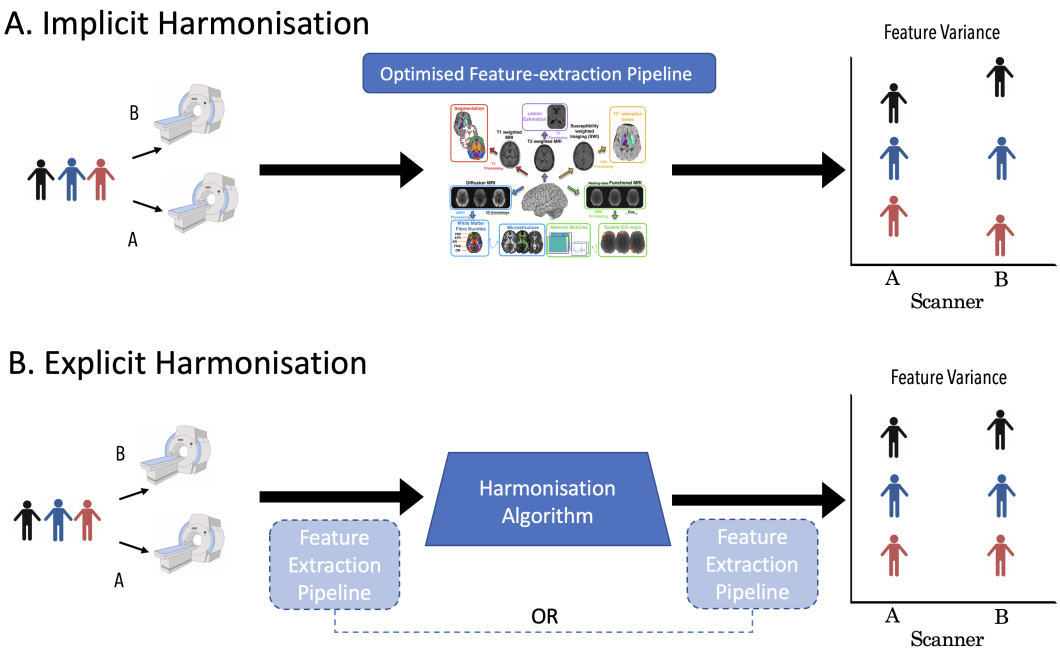

Figure 5.1: *Outlining the general principles of implicit and explicit harmonisation. Explicit harmonisation approaches can be applied either to the data directly or to extracted features, i.e. either before or after any feature extraction pipelines. We anticipate that between-scanner variability is reduced with implicit harmonisation, but can be minimised with explicit harmonisation approaches.*

In this chapter, we demonstrate how our database can be used as a testbed to evaluate and optimise both implicit and explicit harmonisation approaches. Firstly, we use it to optimise pipelines for structural, diffusion and functional imaging data. For structural data we assess the reproducibility of extracted volumes of cortical areas from anatomical images, using a purely atlas-based approach compared to FreeSurfer, that combines registration with an atlas but also within-subject landmarks for areal segmentation. Similarly we evaluate reproducibility of extracted volumes of subcortical areas using unimodal vs multi-modal segmentation approaches. For dMRI data, we compare the reproducibility of ROI-wise DTI based metrics in white matter, when ROIs are defined using a white matter atlas compared to ROIs obtained using subject-specific tractography. We use our data to assess the impact of denoising on diffusion data and functional data. Our hypothesis is that in general, results obtained from data which have been properly denoised will have a higher degree of consistency and between-scanner reproducibility compared to those obtained from undenoised data.

Secondly, we use our data to assess the efficacy of different explicit harmonisation methods, in particular ComBat (Fortin et al. 2017), as one of the most typically used harmonisation methods and the more recent, machine learning-based Neuroharmony (Garcia-Dias et al. 2020). We do this firstly by comparing them against each other and secondly, by assessing the performance of one the methods on different imaging modalities. We finally explore, for both implicit and explicit harmonisation methods their effect on the between-scanner consistency of cross-subject ranking.

5.2 Methods

5.2.1 Implicit Harmonisation

As explained before, we used our data to assess which combinations of data processing and feature extraction steps are optimal for given imaging modalities and features. A pipeline is assumed to be optimal if the features it produces demonstrate minimum between-scanner variability for the scans of the same individual.

5.2.1.1 Cortical areal volumes extracted from T1w data

We assessed how the between-scanner reproducibility of cortical area volumes varied depending on the pipeline used. We compared two options, cortical areas/regions of interest (ROIs) obtained using an atlas-based registration method and FreeSurfer (Desikan et al. 2006, Fischl et al. 2004).

For the registration-based method, which is used in the UK Biobank pipeline (Alfaro-Almagro et al. 2018), the Harvard-Oxford cortical atlas (Makris et al. 2006) was used to warp cortical areas into subject space. The T1-weighted image of a subject was skull-stripped using FSL's Brain Extraction Tool (BET) (Smith 2002). Following brain extraction, the image then underwent a non-linear registration from native T1 space into standard MNI space using FMRIB's Nonlinear Image Registration Tool (FNIRT) (Andersson et al. 2007). The cortical labels from the Harvard-Oxford atlas were warped from MNI space to subject native space to define 97 distinct ROIs. Each of these ROIs were further restricted into voxels labelled as grey matter using FAST (Zhang et al. 2001)) tissue segmentation.

For the FreeSurfer derived metrics, the T1-weighted images were processed using the recon-all function from FreeSurfer version 7.1.0 using the default

settings. The same random seed was used for all runs to avoid run-rerun compute variability between subjects. FreeSurfer areal volumes were extracted for two parcellations, one coarser (66 ROIs based on the Desikan-Killiany (DK) atlas) (Desikan et al. 2006) and one finer (148 ROIs based on the Destrieux atlas (Destrieux et al. 2010)).

For all extracted cortical areal volumes, we calculated the coefficient of variation for each ROI obtained from the same subject scanned in the same scanner and compared these values to those obtained from between-scanner repeats. We also compared the effects of each approach on the consistency of cross-subject ranking. The method for calculating the cross subject ranking is the same as that outlined in section 4.3.5.

5.2.1.2 Subcortical volumes extracted from anatomical MRI data

We performed a similar assessment on the reproducibility of subcortical ROI volumes extracted using different segmentation tasks. We calculated the reproducibility of ROI volumes using imaging data from a single modality and compared this to volumes derived from multi-modal data. The approach for each method is outlined below.

For unimodal segmentation, we used FMRIB's Integrated Registration and Segmentation Tool (FIRST) (Patenaude et al. 2011). The run_first_all function was used which operates with the setting tuned (the number of nodes and boundary conditions) to be optimal for each structure. FIRST was used to segment T1-weighted images into 10 sub-cortical structures. This was run after the data had undergone bias field inhomogeneity correction. We also used FreeSurfer to extract subcortical volumes. Metrics were extracted from a common set of sub-cortical structures so the comparison between the methods would be fair. These were the following (left and right side): Thalamus,

Putamen, Palliduium, Hippocampus, Amygdala

The multi-modal segmentation was performed using FSL's Multimodal Image Segmentation Tool (MIST) (Visser et al. 2016). MIST can use complementary information in different MRI modalities and can therefore be more robust in cases where the contrast in a single modality is not good enough. MIST is trained to learn about the appearance of a structure in a particular set of images, we compared 2 cases of running MIST 1) Training using all 10 subjects from all 6 scanners (i.e. 60 sessions) 2) Training individually per scanner (i.e 10 sessions per scanner) . Multi-modal data for each subject were aligned with that subject's T1-weighted image using a linear rigid body transformation for T2-weighted data and using boundary-based registration (BBR (Greve & Fischl 2009)) for dMRI data.

We assessed the reproducibility of subcortical volumes extracted from MIST in two ways: 1) we use the T1-weighted image and the T2-weighted-FLAIR image of each subject to inform the segmentations 2) we use the T1-weighted image, the T2-weighted-FLAIR image and the FA image from diffusion data to inform the segmentations. It is expected that volumes derived with multimodal information will be more reproducible than those derived with unimodal information. We assessed the extent to which this is true by comparing the coefficient of variation of both methods and the consistency of cross-subject ranking.

5.2.1.3 Tract-wise DTI microstructure metrics

For dMRI data, we assessed the reproducibility of DTI FA measures averaged over white matter ROIs. The ROIs were defined by a skeletonised atlas, by tractography and by skeletonised tractography. The approach for each method is outlined below.

For the skeletonised atlas approach, ROIs in white matter were obtained from the Johns Hopkins (JHU) atlas (Mori et al. 2005, Wakana et al. 2007), which contains 48 ROIs in standard space. These ROIs were further constrained using a TBSS WM skeleton, which effectively depicts the main core of white matter (i.e. regions of WM with the highest FA values)). Using the TBSS pipeline (Smith et al. 2006), all FA data were non-linearly aligned into standard space. Each subject's FA data was then projected onto the TBSS skeleton in standard space to create a skeletonised version of each subject's FA. The skeletonised FA of each subject was then averaged within each JHU ROI.

We defined white matter ROIs using subject-specific tractography. The XTRACT (Warrington et al. 2020) tool was used with default settings which stores tractography results for each individual subject in standard space. XTRACT requires as an input crossing fibres data from the BEDPOSTX (Behrens et al. 2007, Jbabdi et al. 2012, Hernández et al. 2013) tool. Using the BEDPOSTX output, probabilistic tractography was then performed using probtrackx2 (Behrens et al. 2007, Hernandez-Fernandez et al. 2019) called within XTRACT. This generated 42 WM tracts for each subject (after thresholding the paths distribution for each tract at 0.1%), within each of which the mean FA was obtained.

We note that a direct and unbiased comparison between these two methods is a challenge to achieve as the approach of defining ROIs with a skeletonised atlas inherently has an advantage since the metrics are always calculated within regions with minimal partial volume (i.e. within the main core of each tract). Therefore, we subsequently skeletonised the XTRACT-obtained ROIs and computed the mean FA over these skeletonised tractography masks. Furthermore, rather than considering the full set of ROIs we only considered those that were common between the two methods which reduced the number to 9 tracts. For each method, we computed the distributions of coefficients of vari-

ation the consistency in cross-subject ranking.

5.2.1.4 Denoising resting-state functional MRI data

Resting-state fMRI data are typically contaminated by many sources of non-physiological fluctuations. For instance, the temporal evolution of the signal may reflect effects of motion, non-neuronal physiology, scanner artefacts and other nuisance sources (Salimi-Khorshidi et al. 2014). While fMRI denoising methods have been shown to qualitatively improve image quality and robustness of extracted functional features, what is less explored is the effects of denoising on achieving consistent and reproducible results across within- and between-scanner repeats.

We therefore compared two denoising rfMRI methods, one supervised and another unsupervised. These are approaches based on Independent Component Analysis (ICA), which automatically remove components classified as noise after the data have been decomposed.

The supervised denoising methods we used was FIX (Beckmann & Smith 2004, Griffanti et al. 2014, Salimi-Khorshidi et al. 2014). We ran v1.06 which with the optional flags -m (which cleans up motion confounds) and the highpass value, -h, was set to 100. The running of FIX requires training data and in our case this consisted of 40 UK Biobank rfMRI datasets (i.e. acquired using a Siemens scanner) with the noisy components labelled by hand. We compared this to ICA-AROMA (Pruim, Mennes, van Rooij, Llera, Buitelaar & Beckmann 2015) which is an unsupervised method. We run ICA-AROMA v0.3 beta after spatial smoothing, but prior to temporal filtering within the fMRI preprocessing pipeline. We compared both of these with each other and also with respect to no denoising. For each of these instances we computed the distributions of coefficients of variation for the node amplitudes and the

consistency in cross-subject ranking.

5.2.1.5 Evaluating the effects of diffusion MRI denoising

Diffusion MRI denoising methods aim to remove thermal noise effects. This differs from the noise present in rfMRI which reflects multiple sources of structured noise such as motion. The data that we have acquired allows us to evaluate the efficacy of dMRI denoising methods by assessing the impact of denoising on the reproducibility of results obtained from within-scanner repeats. If the majority of within-scanner repeat variability is driven by thermal noise, then a denoising technique is expected to reduce this variability significantly.

To assess this, we denoised the diffusion data using Marchenko-Pastur Principal Component Analysis (MP-PCA) (Veraart et al. 2016). We used dwidenoise which is part of of the MRtrix3 package v 3.0.3. The denoising was performed on the raw diffusion data prior to any movement or distortion corrections. Following denoising, the data was fed through the diffusion pipeline described in section 4.2.5 and IDPs were extracted. We performed this on the subjects for which we had acquired within-scanner repeats and compared the reproducibility of the results with those obtained from undenoised data.

5.2.2 Explicit Harmonisation

In this section we assess the effectiveness of existing harmonisation approaches. Any harmonisation approach that claims to explicitly remove site effects should inherently reduce between-scanner variability in imaging features compared to not using harmonisation at all. We evaluate the performance of each method by comparing the resulting variability with that of within-scanner variability

which we use as as a reference/benchmark.

We evaluated the performance of two approaches that harmonise directly imaging-derived features. The first approach is ComBat (Fortin et al. 2017), which relies on the presence of a representative dataset and harmonises entire cohorts for any feature that can be derived from the available data. The second is Neuroharmony (Garcia-Dias et al. 2020) a supervised machine learning approach, which can harmonise individual datasets only for features it has been trained on. These two approaches have been chosen because they harmonise a common set of features which enables a fair comparison between them.

We first tested and compared these methods for features that Neuroharmony has been trained on. These include the volumes of 101 cortical (DK atlas (Desikan et al. 2006)) and subcortical (ASEG atlas (Fischl et al. 2002)) ROIs derived using the recon-all function from FreeSurfer. We ran FreeSurfer on all scanning sessions for which we had within-scanner repeats. We evaluated the performance of these methods by comparing them to the variability of the unharmonised values and using the variance of the within-scanner repeats as a baseline.

Neuroharmony for brain ROI volumes

Neuroharmony harmonises data on a subject by subject basis using a model trained to map quality metrics for structural images, extracted from MRIQC (Esteban et al. 2017), to harmonised values that have been obtained from ComBat. We run MRIQC version 0.16.1 to obtain 68 Image quality metrics (IQMs) which we used as inputs for each subject to the pre-trained Neuroharmony model (trained on 15,026 subjects). For each subject, we used a 71-element vector (68 IQMs + age, sex and the original relative volumes of the ROIs) as input variables to predict the ComBat corrections for each ROI. For each session, the volumes of the 101 ROIs were normalised by dividing the

volume of each region by the total intracranial volume of the subject, as it was done in (Garcia-Dias et al. 2020).

ComBat for brain ROI volumes

We run ComBat version 0.2.12. Our input data was a matrix of dimensions 24 (4 subjects $\times$ 6 scanners) $\times$ 101 (number of FreeSurfer ROI's) (We note that ComBat was also ran for the full cohort of 10 subjects, but trends were very similar to the ones obtained using a subset of 4 subjects that had within-scanner repeats). For the batch variables (the variables which encode for the scanner) we had 6 unique variables each representing one of the six scanners from which the data were acquired. For the categorical variable to be preserved during harmonisation, we specified gender and for the continuous variable to be preserved, we specified age. Combat was run with default settings in which case, empirical Bayes is turned on (information is pooled across features rather than the harmonisation model being fit for each feature separately) and parametric adjustment is performed.

ComBat for other features

Similar to running Combat for the brain ROI volumes, we assessed its efficacy in harmonising features of other types or from other modalities. Specifically we applied Combat in subcortical volumes obtained from FIRST, in the T2* values of subcortical regions extracted from susceptibility-weighted images and in the FA of white matter ROIs obtained from dMRI. These IDPs are chosen as they are standard outputs from the UK Biobank pipeline.

5.3 Results

The results we show demonstrate how implicit and explicit harmonisation approaches perform when compared when using within-scanner variability as a benchmark. For this reason, the results of subject averages comprise only the

4 subjects for which within-scan repeats were obtained.

5.3.1 Implicit Harmonisation

5.3.1.1 Variability of cortical area volumes

We first explored how different approaches for obtaining cortical area volumes affected the within-subject variability of these IDPs across-scanners. We used the variability of within-scanner repeats for the same subjects as a baseline reference. Figure 5.2 shows that for images acquired within the same scanner, cortical volumes derived using an atlas-based registration method have a smaller baseline variability than those derived using FreeSurfer. However, when looking at variability of IDPs for data acquired on different scanners, we see that on average, ROI volumes derived using FreeSurfer (both coarse and fine parcellations) are more robust than the simpler registration-based approach. The results also show that for FreeSurfer-extracted ROIs, fine parcellations are more variable than coarse parcellations, which is expected as the finer parcellations lead to smaller regions. Nevertheless, even if the number of regions for Registration-based is smaller than the fine FreeSurfer parcellation approach (97 vs 148 ROIs respectively), FreeSurfer volumetric measures seem more robust to between-scanner effects. We also show results for a sub-analysis where we extracted the metrics for the common ROIs across the an atlas-based registration and coarse parcellations from FreeSurfer. These results are shown in Figure 5.3. The common structures are (left and right sided): parahippocampal, postcentral, superiorfrontal, frontalpole and insula.

When looking into preservation of cross-subject ranking across scanning sessions, both approaches show relative high correlation of cross-subject ranking yet the correlation of subject ranking for FreeSurfer extracted volumes is better than the registration-based approach, for both fine and coarse parcellations.

Table 5.1 shows for each approach, the percentage difference of the values obtained from images acquired within the same scanner compared to within different scanners. The results show that on average, the approach with the highest increase in the median CoV values across the ROIs was for registration-based methods. Both approaches from FreeSurfer show percentage differences which are almost the same and about 4 times less than the increase in variability compared to the registration-based approach.

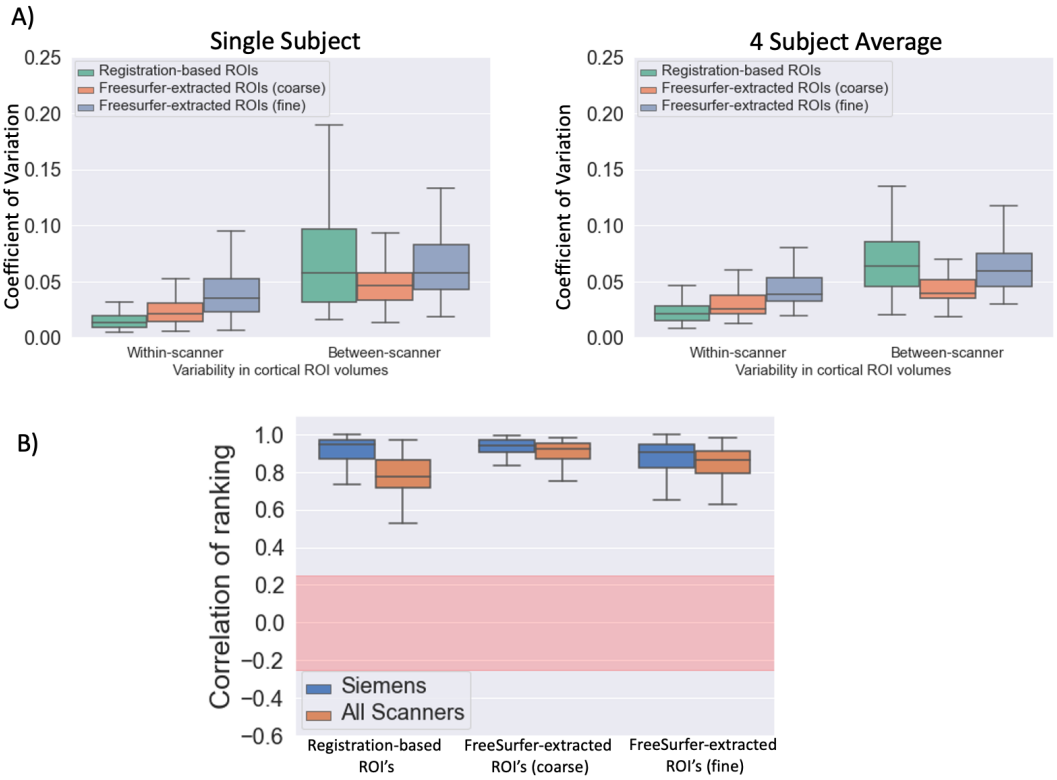

Figure 5.2: *Reproducibility of ROI-wise cortical volumes A) Distribution of CoVs for Registration based ROIs (97 ROIs) compared to FreeSurfer extracted ROIs (coarse - 66 ROIs and fine - 148 ROIs) B) Correlation of cross subject ranking for Registration-based ROIs compared to FreeSurfer-extracted ROIs (coarse and fine).*

Table 5.1: Percentage difference of median between-scanner CoV of cortical ROI volumes with respect to (w.r.t.) the within-scanner CoV.

Approach	Increase of between-scanner CoV w.r.t. within-scanner CoV
Registration-based ROIs	206%
FreeSurfer-extracted ROIs (Coarse)	52.7%
FreeSurfer-extracted ROIs (Fine)	52.1%

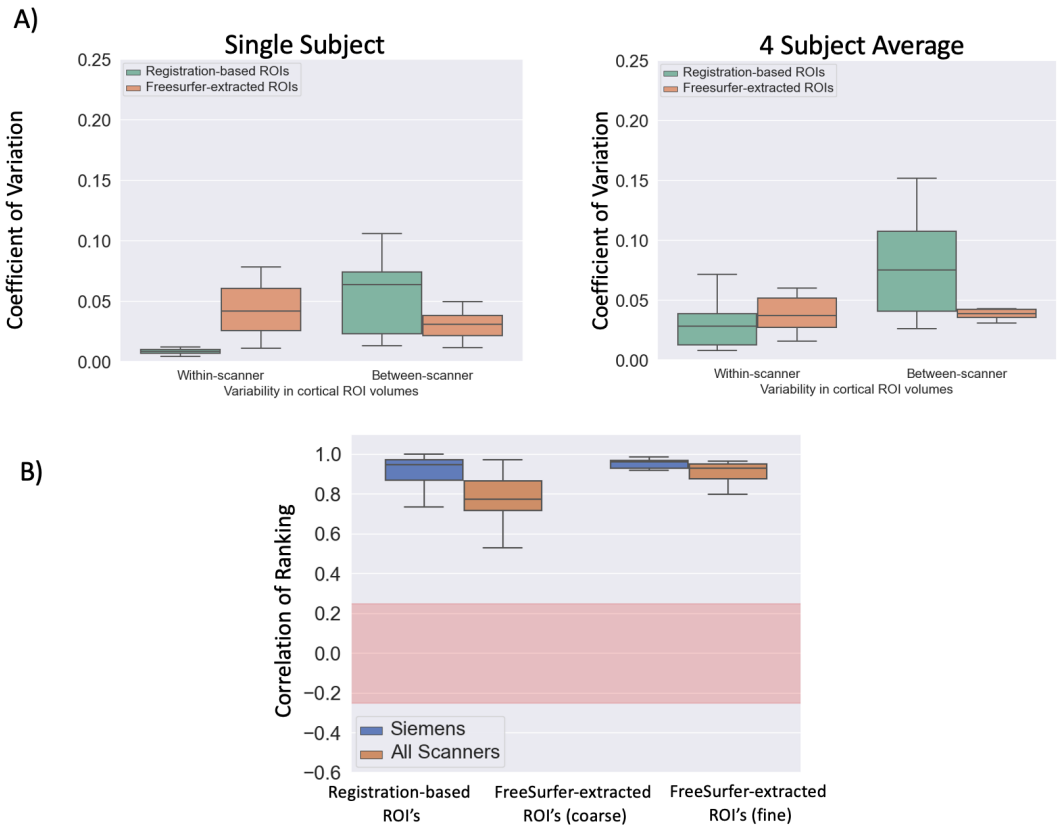

Figure 5.3: *Reproducibility of ROI-wise cortical volumes A) Distribution of CoVs for Registration based ROIs compared to FreeSurfer (coarse) extracted ROIs. Metrics shown are from a subset of 10 ROIs which are common to both methods. B) Correlation of cross subject ranking for Registration-based ROIs compared to FreeSurfer-extracted ROIs (coarse).*

5.3.1.2 Variability of subcortical volumes

We compared the reproducibility of ROI-wise subcortical volumes derived using a range of segmentation algorithms, specifically unimodal and multi-modal segmentation. The results in Figure 5.4A show that, on average, sub-cortical volumes derived using multimodal segmentation approaches are less variable than those relying on only one modality for segmentation. This was true regardless of implementation details for the multimodal approach (number of modalities and number of training sessions considered).

For multi-modal segmentation, our results showed that training MIST on all scanning sessions was not considerably different in reducing between-scanner variability compared to training it on many sessions from an individual scanner.

Multimodal segmentation using data from 2 anatomical modalities resulted in less variable segmentations than training it on 3 modalities (anatomical and diffusion) in absolute terms. However, Table 5.2 shows that training MIST on 3 modalities was more successful at preserving the level of variability similar to that in the within-scanner repeats and in that respect was the more successful approach.

Figure 5.4B shows that all approaches show a high level of correlation of ranking across subjects. (at least 80%) in all cases. In agreement with the previous trends, there is a benefit in preserving between-scanner cross-subject rankings with multimodal segmentations.

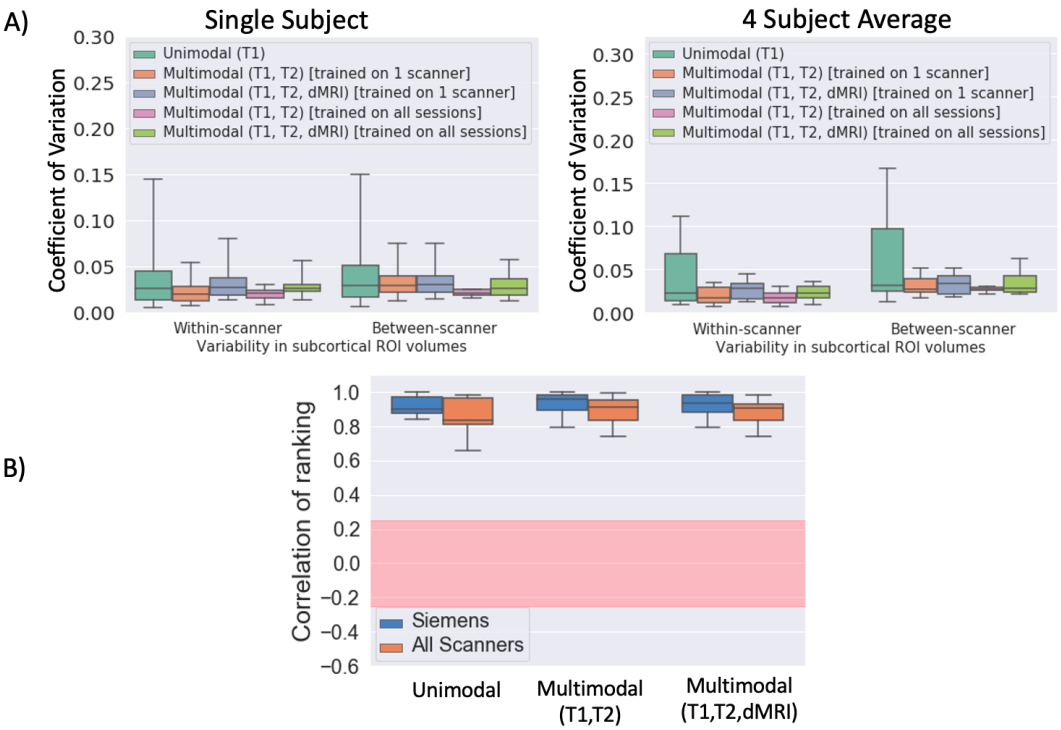

Figure 5.4: *Reproducibility of ROI-wise subcortical volumes A) Distribution of CoVs for Segmentation-based ROIs using one modality, two modalities and three modalities B) Correlation of cross subject ranking for Registration based Segmentation-based ROIs using one modality, two modalities and three modalities.*

Table 5.2: Percentage difference of median between-scanner CoV of sub-cortical ROI volumes with respect to (w.r.t.) the within-scanner CoV.

Approach	Increase of between-scanner CoV w.r.t. within-scanner CoV
Unimodal	42.7%
Multimodal (T1,T2) Trained on 1 scanner	62.0%
Multimodal (T1,T2,dMRI) Trained on 1 scanner	19.5%
Multimodal (T1,T2) Trained on all sessions	59.1%
Multimodal (T1,T2,dMRI) Trained on all sessions	30.1%

5.3.1.3 Variability of Tract-wise DTI FA

Figure 5.5A shows that, on average, DTI FA values, averaged over white matter ROIs, are more reproducible between-scanners when the ROIs are obtained from a skeletonised atlas compared to ROIs obtained from subject-specific tractography. This is true regardless of whether the tractography ROIs are skeletonised or not.

Table 5.3 shows for each approach, the percentage difference of the median CoV of ROI-wise FA values obtained from between-scanner vs within-scanner data. The results confirm the above trends and show the advantage of skeletonising the XTRACT-obtained ROIs.

Figure 5.5B shows that for all sessions there is greater consistency in cross-subject ranking for ROIs obtained from a skeletonised white matter atlas compared to tractography regardless of whether the ROIs were skeletonised. However, it is important to note that just as skeletonising the ROIs increased between-scanner reproducibility in Figure 5.5A, doing this also increased consistency in between-scanner subject rankings.

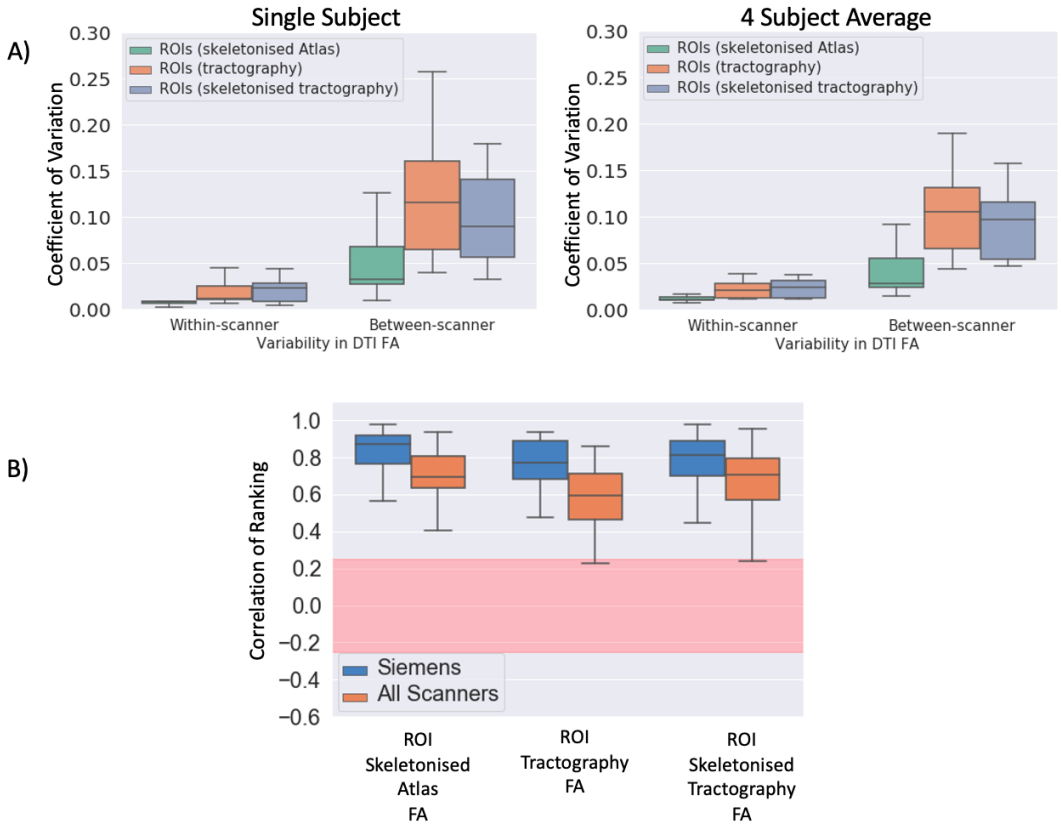

Figure 5.5: *Reproducibility of ROI-wise FA values A) Distribution of coefficients of variation for ROIs obtained from a skeletonised atlas, tractography and skeletonised tractography B) Correlation of cross subject ranking for ROIs obtained from a skeletonised atlas, tractography and skeletonised tractography.*

Table 5.3: Percentage difference of median between-scanner CoV of tract-wise FA with respect to (w.r.t.) the within-scanner CoV.

Approach	Increase of between-scanner CoV w.r.t. within-scanner CoV
ROIs Skeletonised Atlas	159%
ROIs (Tractography)	407%
ROIs (Skeletonised Tractography)	297%

5.3.1.4 Variability of rfMRI node amplitudes

We compared the effect of different denoising approaches on the reproducibility of rfMRI node amplitudes. As FIX is a supervised denoising approach that has been trained on Siemens data, we anticipate that it works better for Siemens rather than non-Siemens data. We therefore looked separately into subsets of data and Figure 5.6A shows 2 exemplar subjects: one for which within-scan repeats have been acquired on a Phillips Achieva scanner (Subject A) and the other for which within-scan repeats have been acquired on a Siemens Prisma

scanner (Subject B). Results show that for within-scanner repeats acquired on a Siemens scanner, FIX denoising is more effective at reducing variability than ICA-AROMA (i.e. a supervised method trained on Siemens data is better than an unsupervised method), whereas for within-scanner repeats acquired on a Phillips scanner, the unsupervised ICA-AROMA denoising is more effective than FIX. When looking into between-scanner variability, both denoising approaches reduce variability compared to raw data, but the unsupervised ICA-AROMA performs better than FIX.

Figure Figure 5.6B shows that the very low between-scanner correlation of cross-subject ranking of the node amplitudes is not improved and stays low even after denoising for both approaches. This reflects the magnitude of the challenge associated with harmonising rfMRI data.

5.3.1.5 The effect of dMRI denoising

Figure 5.7 shows the percentage difference in coefficients of variation of tract-wise FA values, before and after denoising averaged across 4 subjects. Our hypothesis was that denoising the data would yield more consistent results across within-scanner repeats. However, on average, this is true for only 28/48 (58.3%) of the considered IDPs. For a number of IDPs, denoising caused the opposite effect of what was anticipated, resulting to more variability between repeats. Figure 5.7 shows the major outliers (a difference of greater than 15%) and what we note is that these are related to tracts at frontal/inferior parts of the brain. The specific regions highlighted are the left and right cerebellar peduncle and the right uncinate fasciculus.

To investigate this further, we investigated the corresponding white matter ROIs in the denoised and the undenoised data. This comparison is shown in Figure 5.8. We show the FA maps in standard space derived from denoised

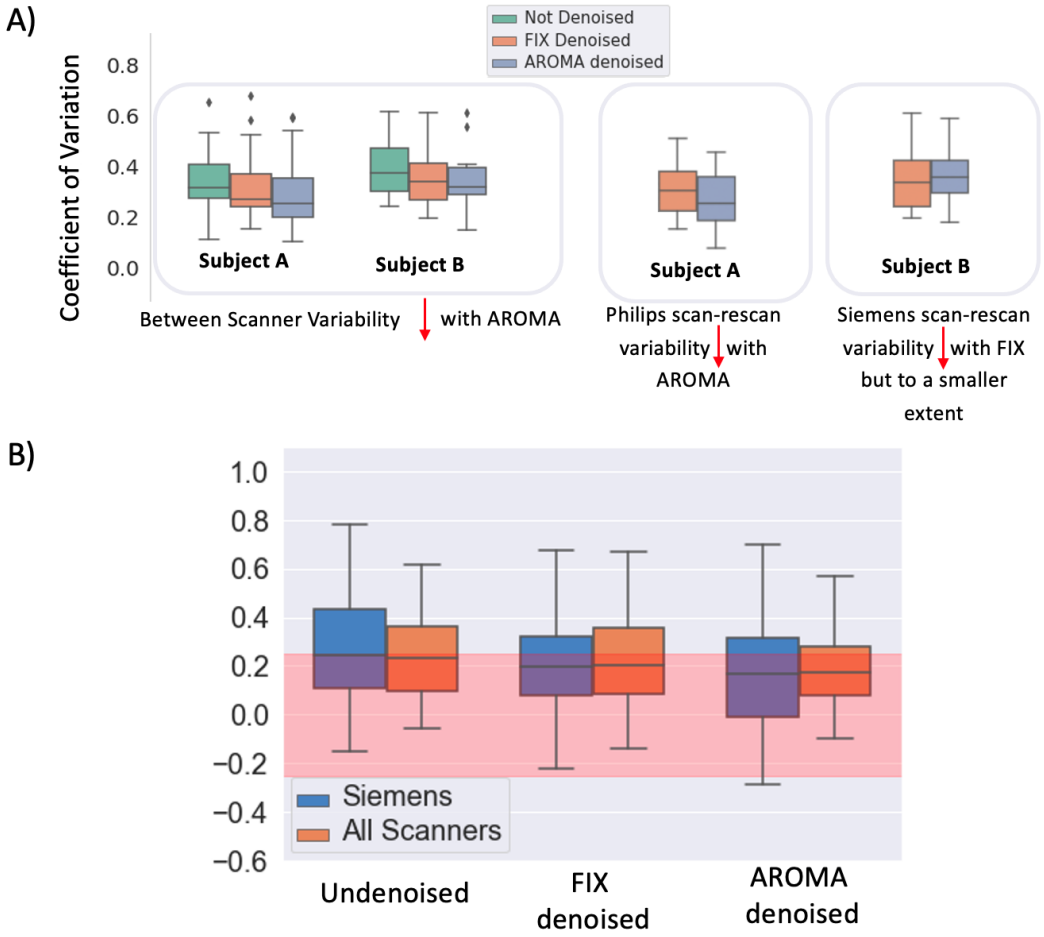

Figure 5.6: *Reproducibility of rfMRI Node Amplitudes based on method of denoising. A) Distribution of CoVs across between-scanner repeats for node amplitudes comparing raw data with denoising with FIX (supervised) and ICA-AROMA (unsupervised) for 2 subjects (leftmost panel). Coefficient of variation after denoising for within-scanner repeats for a subject acquired on a Philips scanner (central panel) and on a Siemens scanner (rightmost panel). B) Correlation of cross-subject ranking for the denoising approaches.*

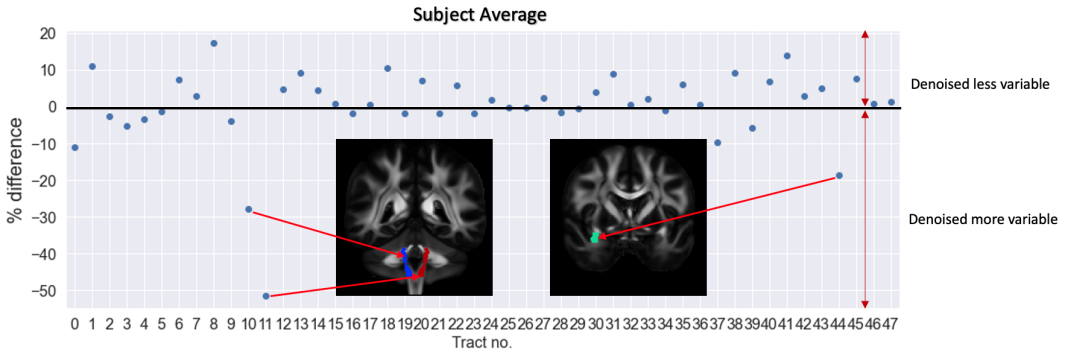

Figure 5.7: *The average across 4 subjects of the difference in within-scan repeat coefficients of variation of ROI-wise DTI FA before and after dMRI denoising.*

and undenoised dMRI data with the left and right cerebellar peduncle highlighted. We see that the denoised FA maps in the highlighted regions deviate

significantly from the region enclosed by the region of interest taken from the JHU atlas (Mori et al. 2005, Wakana et al. 2007). In the undenoised version, this is not the case and we see that most of the tissue lies within the ROI, explaining why FA values from these regions appeared different in the previous plot. As the main outliers were found in ROIs close to areas with large susceptibility-induced distortions, the interplay between dMRI (patch-based as done here) denoising and distortion correction may be a factor contributing to these observations.

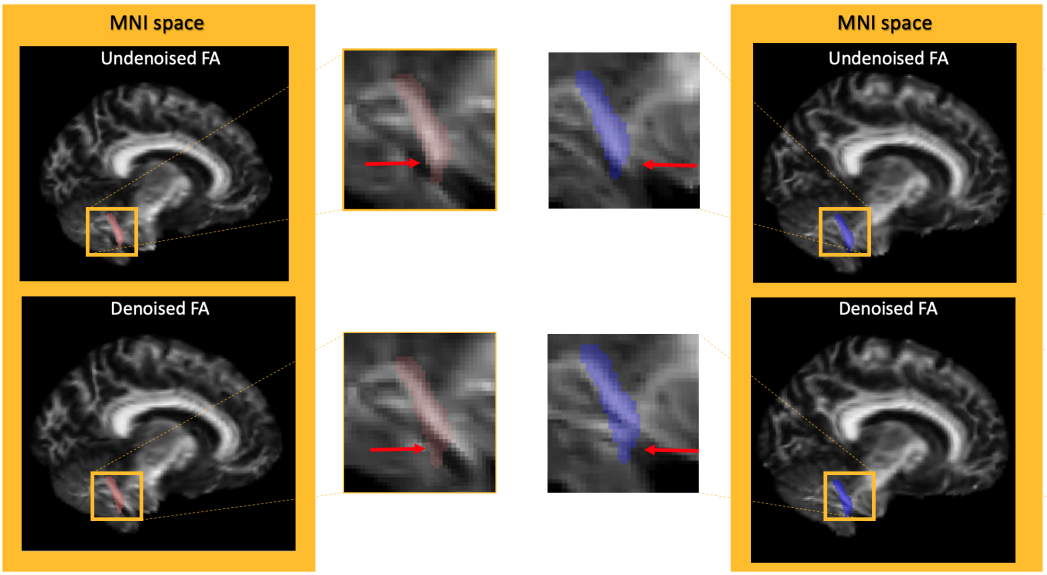

Figure 5.8: *FA maps of one subject derived from denoised and undenoised dMRI data with the left and right cerebellar peduncle highlighted in MNI standard space. These tracts correspond to the outliers identified in Figure 5.7.*

5.3.2 Explicit Harmonisation

In addition to identifying optimal processing steps for pipelines to minimise between-scanner variability in imaging extracted features (implicit harmonisation), we used the data to evaluate existing harmonisation approaches, using the within-scanner variability as a baseline. These are meant to explicitly reduce between-scanner variability. Figure 5.9 shows how between-scanner variability from harmonised cortical volumes from Neuroharmony and Com-

Bat compare to variability in unharmonised data and within-scanner repeats for the same subjects. Both harmonisation approaches reduce between-scanner variability (median CoV=0.074) in the considered features, with ComBat (median CoV=0.051) reducing it more than Neuroharmony (median CoV=0.073). Even so, they are still higher than within-scanner variability of the same features (median CoV=0.027).

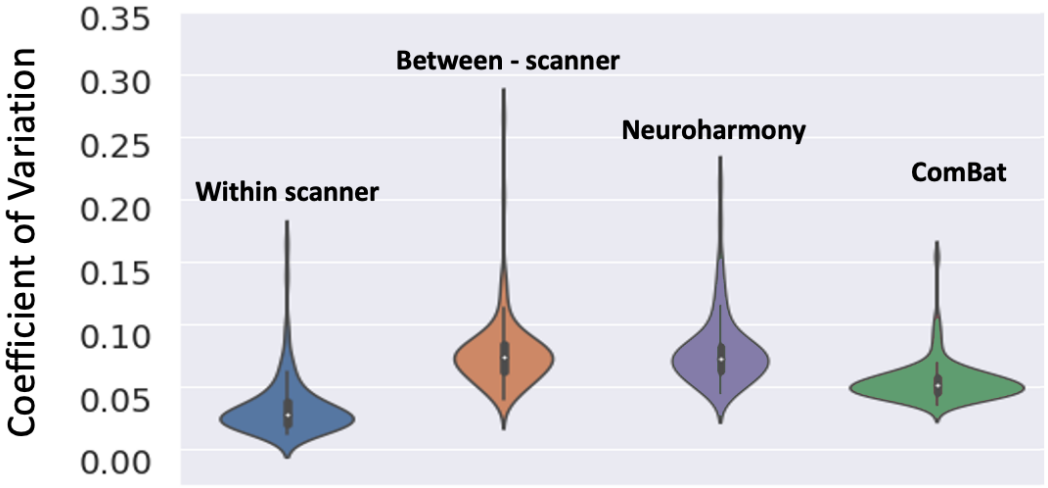

Figure 5.9: *The effect of harmonising cortical area volumes using ComBat and Neuroharmony for 4 subjects. For each subject and each imaging feature, a CoV was computed against 6 repeats (either within-scanner or between-scanner), prior to harmonisation. Violin plots are made from the CoVs of all considered features. After harmonisation, the CoVs depict the harmonised between-scanner repeats.*

We further explored the effects of explicit harmonisation approaches for other features. As Neuroharmony is a supervised model trained only for a certain feature (cortical area volumes), we used only ComBat for different features, including subcortical volumes obtained from FIRST, the T2* values extracted from susceptibility-weighted images and the FA of white matter ROIs obtained from diffusion MRI. These were once again obtained from a common set of tracts shared by the two approaches.

Figures 5.10, 5.11 and 5.12 present the results. For all features we see that on average, ComBat reduces the between-scanner variability although this variability is not always as low as within-scanner variability. Combat seems to be

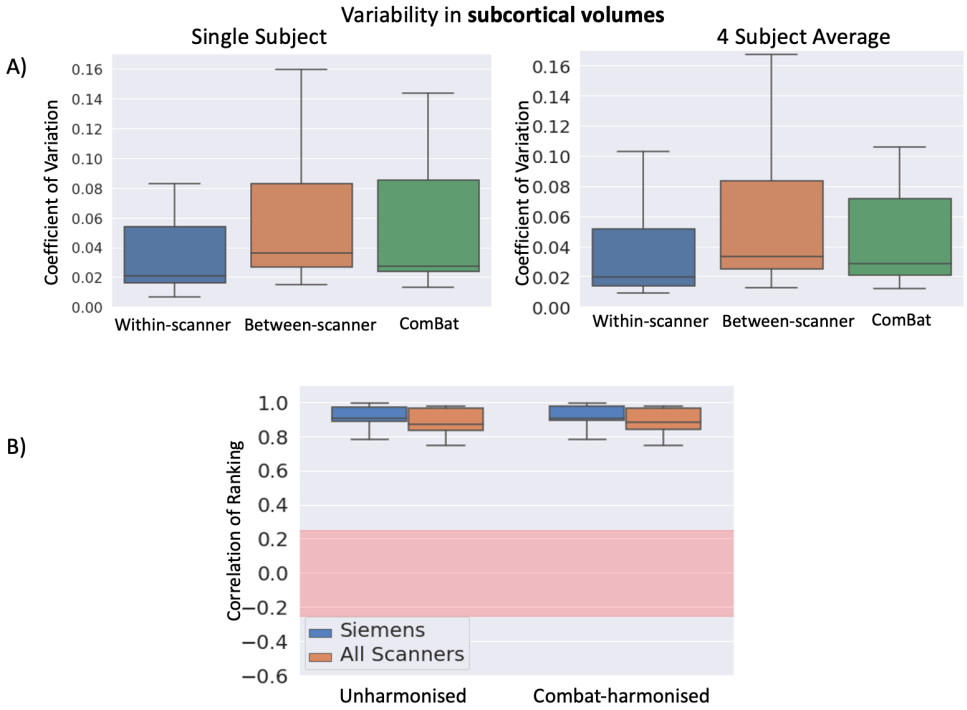

Figure 5.10: *A) The effect of harmonising subcortical volumes from multi-site data with ComBat. For each subject and each imaging feature, a CoV is computed against 6 repeats (either within-scanner or between-scanner), prior to harmonisation. Box plots are made from the CoVs of all considered features. After harmonisation, the CoVs depict the harmonised between-scanner repeats. B) Correlations in between-scanner cross-subject ranking for each measure with and without ComBat harmonisation.*

more effective in reducing between-scanner variability for some features (e.g. FA or T2*) compared to others (e.g. subcortical volumes). Interestingly, a common trend across all features is that between-scanner cross-subject rankings before and after harmonisation are almost identical. Combat modifies feature values such that variability is reduced, but it is not beneficial in restoring cross-subject ranking between scanners.

5.4 Discussion

We have used our dataset as a testbed to explore and evaluate harmonisation approaches. Specifically, we have used our data to identify various optimal processing steps used in feature extraction pipelines, such that between-scanner variability in extracted features is minimised compared to e.g. within-

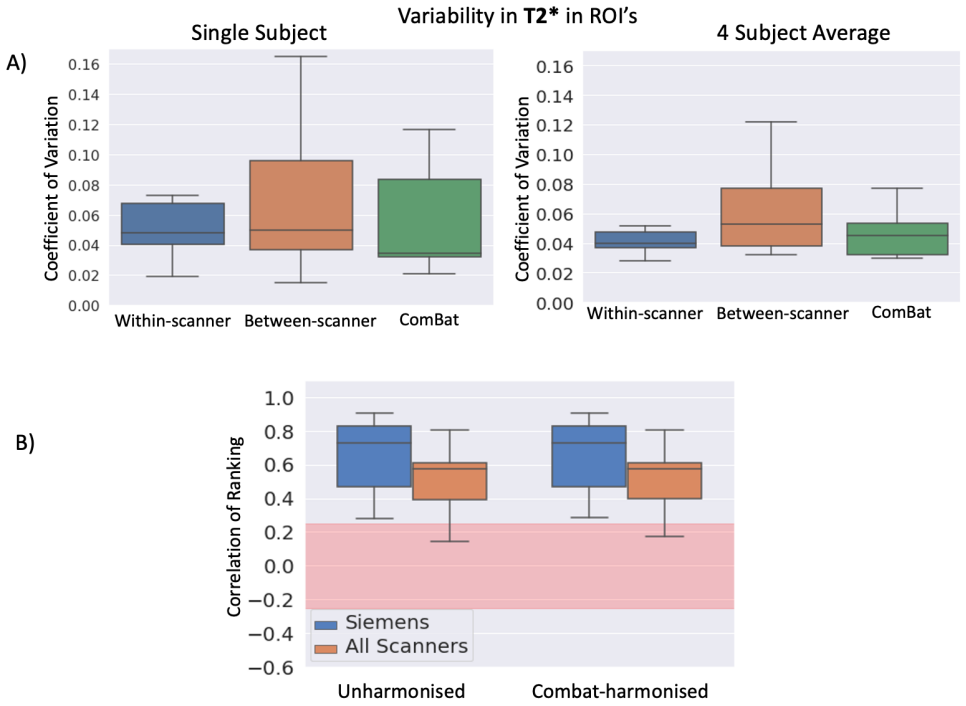

Figure 5.11: *A) The effect of harmonising T2* values in subcortical ROIs from multi-site data with ComBat. For each subject and each imaging feature, a CoV is computed against 6 repeats (either within-scanner or between-scanner), prior to harmonisation. Box plots are made from the CoVs of all considered features. After harmonisation, the CoVs depict the harmonised between-scanner repeats. B) Correlations in between-scanner cross-subject ranking for each measure with and without ComBat harmonisation.*

scanner variability (implicit harmonisation). We have also tested performance of post-processing harmonisation tools (explicit harmonisation) and specifically checked whether the harmonised features between-scanners are indeed less variable (and by how much) compared to no harmonisation. For these tests we also used our data to establish within-scanner variability baselines for the harmonised features.

For anatomical imaging features, we have found a number of interesting trends. Cortical area volumes extracted from FreeSurfer and subcortical volumes extracted from multi-modal segmentation seem to have less between-scanner variability compared to other approaches explored. Previous studies have shown that cortical volumes derived from FreeSurfer have a strong degree of robustness against scanner effects. For instance in (Iscan et al. 2015) it is shown

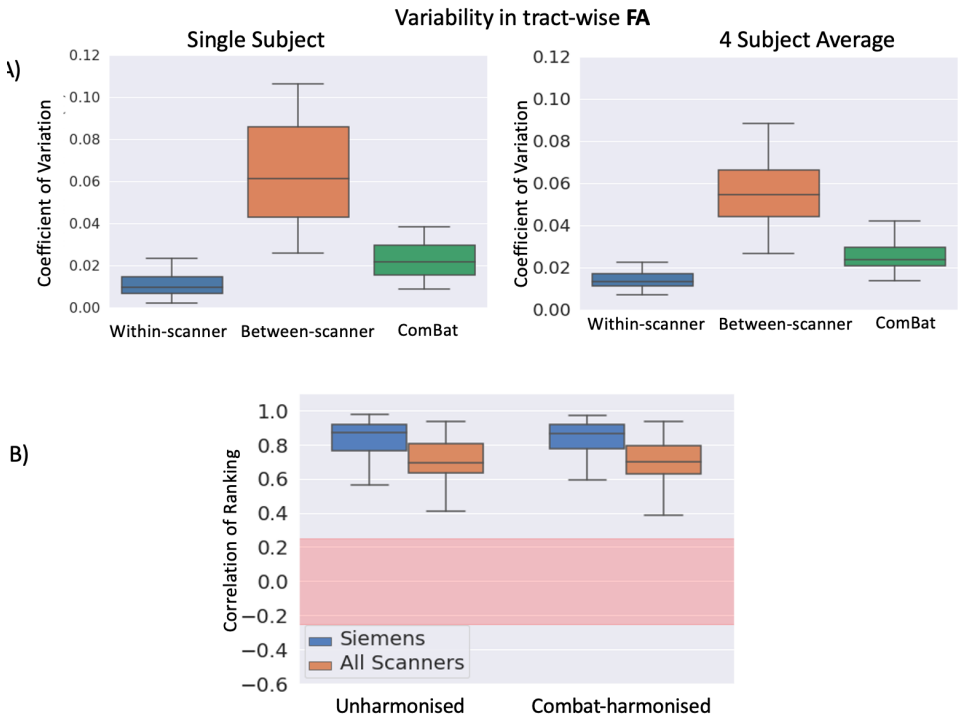

Figure 5.12: *A) The effect of harmonising tract-wise FA values from multi-site data with ComBat. For each subject and each imaging feature, a CoV is computed against 6 repeats (either within-scanner or between-scanner), prior to harmonisation. Box plots are made from the CoVs of all considered features. After harmonisation, the CoVs depict the harmonised between-scanner repeats. B) Correlations in between-scanner cross-subject ranking for each measure with and without ComBat harmonisation.*

that for the DK atlas, cortical volume measures showed test-retest correlations (scans acquired at 4 different sites) scores of 0.88. This study also showed higher test-retest correlation and inter-class correlation scores for volumes for the DK atlas (coarse) than the Destrieux atlas (fine) which is in agreement with the results we obtained. These results confirm what we expect since regions defined by the DK atlas are larger than those in the Destrieux atlas.

For subcortical volumes, we found volumes derived using a multi-modal segmentation method (MIST) were more reproducible than those derived using a unimodal approach (FIRST). This is in agreement with the findings in (Visser et al. 2016) who compared the approach with FIRST and FreeSurfer using a manual segmentation as a benchmark. Results showed that the dice overlap was higher for MIST than it was for FIRST indicating that segmentations

performed using multiple imaging modalities are more robust than than those which only used a single modality. We also assessed the advantage of training MIST with data from 3 modalities (T1-weighted,T2-weighted and dMRI data) compared to training it using 2 (T1-weighted and T2-weighted). Intuition would suggest that leveraging imaging information from more modalities would result in more reproducible results, however our results show that adding dMRI data as an input to MIST decreased between-scanner reproducibility. These findings agree with results in (Visser et al. 2016), who found that increasing the number of modalities used for MIST segmentation can increase variability. This can happen for regions where the contrast is very clear from structural images. In this case, segmentations from the structural images alone are highly reproducible and adding another modality, particularly a more noisy one like dMRI, introduces new sources of variability. As shown in the previous chapter the between-scanner variability of dMRI features was considerably more than that of anatomical features. Finally, we found that MIST performs better when trained on all scanning sessions compared to when it has only been trained on one scanner. This confirms intuition as in the former case, data from all scanners is being used to inform the segmentations which is a form of an indirect harmonisation.

Our explorations of between-scanner variability of FA measures in white matter ROIs suggest that FA derived in ROIs defined by a skeletonised atlas are considerably less variable than in ROIs derived using binarised tractography masks, whether skeletonised or otherwise. The increased variability of tractography-defined ROIs compared to skeletonised-atlas-defined ROIs is expected due to the fact that in the former, the sizes of ROIs are allowed to vary whereas in the latter, the sizes of the ROIs are relatively fixed. Previous studies have also assessed the reproducibility of FA measures in ROIs (Vollmar et al. 2010, Heiervang et al. 2006). In (Vollmar et al. 2010), it is reported that methods based on probabilistic tractography can introduce ap-

proximately 50% more variation compared to methods with predefined ROIs (akin to the skeletonised ROIs we have here). Despite these large discrepancies in reproducibility, it is pointed out in (Heiervang et al. 2006) that there is a need to balance reproducibility with sensitivity. Even though a measure can have low between-scanner reproducibility, it may have a higher sensitivity to a relevant changes so we cannot always assess the merit of these methods based on reproducibility alone. We endeavoured to introduce a middle ground in this comparison by restricting the tractography masks within a skeleton. As expected, this increased reproducibility compared to raw tractography results. Whether this maintains the sensitivity we'd hope to preserve would require further investigation.

For resting-state functional MRI data, denoising the data with an unsupervised denoising method was more effective at reducing between-scanner variability of rfMRI node amplitudes, compared to a supervised method trained on data from a specific vendor. Previous comparisons of the two approaches have been conducted (Pruim, Mennes, Buitelaar & Beckmann 2015). ICA-AROMA was shown to have higher levels of resting state network reproducibility than FIX, even when FIX was trained with very high quality data from the Human Connectome Project (Van Essen et al. 2012). Thus we see that in the general case unsupervised denoising may be advantageous across scanners. But when similar training data is available, it may be optimal to use a supervised approach.

We found a slightly unexpected trend for dMRI denoising using MP-PCA (Veraart et al. 2016). Even within-scanner variability of extracted DTI features did not always decrease after denoising compared to features extracted from undenoised data. It is worth pointing out that raw SNR and CNR values do increase after denoising in this data. The natural question to ask is why then does variability of features does not improve after denoising? A possible explanation is that we observed a trend that outlier cases where in caudal regions of

the brain where denoising appeared to have increased variability by more than 15%. These are areas known to be prone to susceptibility artefacts (Andersson et al. 2003) and therefore distortion correction is more impactful in these areas. The fact we see these areas significantly affected after denoising suggests that there is a possible interaction between denoising and distortion correction. This could happen because, even prior to distortion correction, denoising assumes that the every voxel is in the correct place yet this is not true in the presence of distortions. As denoising is patch-based, incorrectly placed voxels would end up influencing the denoising process meaning a distortion correction like this could lead to misplaced voxels and in slightly different ways for the different repeats. This suggests that the optimal way of denoising requires more exploration e.g exploring when in the processing pipeline denoising is more likely to have the desired effects.

We also compared explicit harmonisation approaches in ways that have not been evaluated before. We showed that that both Neuroharmony (Garcia-Dias et al. 2020) and ComBat (Fortin et al. 2017) reduced the between-scanner variability. It is important to note that with 6 scanning sessions, we were at the lower end of the recommended number of subjects for good results using ComBat. In (Maikusa et al. 2021) it is stated that it is not ideal to perform ComBat harmonisation with 20 subjects or less therefore the application of it here was not to its full potential. Nevertheless, ComBat still managed to improve on results obtained from Neuroharmony. As Neuroharmony uses pre-trained corrections provided by ComBat to perform harmonisation on unseen data, it is somewhat expected that it will not perform as well as ComBat which directly uses data from multiple scans within the cohort to harmonise. However, there is a trade-off at play in that while ComBat is more effective at harmonising data, it can not harmonise individual datasets. A limitation of our assessment is that we have been unable to asses the effectiveness of ComBat using the number of subjects recommended in the literature (Maikusa et al. 2021),

which would be necessary in order to assess in optimal operation. We however found small changes when comparing ComBat performance using only the 4 subjects (4 × 6 sessions) that had within-scanner repeats against performance when considering the full cohort of 10 subjects (10 ×6 sessions). As the 20 subjects recommendation was established using 3 between-scanner sessions, it may be that using twice the number of between-scanner sessions in our study makes up for the fewer number of subjects.

Our application of ComBat to other modalities showed that ComBat in general reduced between-scanner variability although it did not reduce it to levels as low as within-scanner variability. We observed that ComBat was more effective at harmonising FA and subcortical T2* features than it was at harmonising subcortical volumes. The relatively small difference in ComBat corrected subcortical volumes to uncorrected volumes is in agreement with findings from other studies (Treit et al. 2022). The authors in the mentioned study used ComBat to reduce systematic variations in brain volumes of 23 travelling subjects scanned in 3 different scanners and they found minimal changes (of less than 5%) between corrected and raw volumes for several sub-cortical regions (caudate, globus pallidus, putamen, and thalamus). The authors in (Treit et al. 2022) point out that the degree to which ComBat decreases inter–subject variability likely depends on the magnitude of site effects in the raw data implying that that ComBat has less of an effect on results which are most robust to site effects. Our findings support this notion as of the three features (subcortical volumes, T2* values and FA values), subcortical volumes had on average the least between-scanner variability of the three and were affected the least by ComBat.

Another important finding for explicit harmonisation approaches is that they did not improve or change inconsistencies in between-scanner cross-subject rankings. This was true across all modalities and features tested. This was

not the case for implicit harmonisation approaches which in most cases had a beneficial effect in that respect, in addition to reducing between-scanner variability.

The work in this chapter enables future opportunities to evaluate a wider range of Harmonisation algorithms and extend upon the two approaches that we considered here. There are significantly more approaches as overviewed in 2 and most of them have been evaluated using ad-hoc criteria. Our data provide a testbed for objective evaluation and comparisons, as we illustrated in this chapter with a number of examples across different imaging modalities and feature types. Furthermore, optimal pipelines for feature extraction can be aimed for, with a good example being that of identifying best ways to denoise dMRI data. We have begun to show how our results an be used to arrive at a consensus of what the most reproducible and generalisable pipelines are to process brain MRI data thus help to address the challenge of poor reproducibility, accuracy and consistency in quantitative MRI. In the final chapter, these results will be summarised in the form of guidelines which will provide recommendations of which tools to use in order to obtain the most consistent and reliable results.

Chapter 6

Summary

6.1 Recommendations and Guidelines

The results of the explorations performed of previous chapter provide the basis for recommendations and guidelines on which tools use to minimise the effects of scanner induced variability. While not all possible pipelines have been explored, this chapter will suggest current best processing tools for use in pipelines.

Volumetric Features

During our exploration of different approaches for obtaining cortical area volumes, we found that ROI volumes derived using FreeSurfer were more robust than those derived using FIRST which is a registration-based approach. Accordingly, we would recommend using FreeSurfer when performing this type of analysis. Specifically, we would recommend using Desikan-Killiany (DK) (Desikan et al. 2006) atlas which has fewer and larger ROIs as these are less susceptible to noise than their finer counterparts from the Destrieux atlas.(Destrieux et al. 2010).

For sub-cortical regions, if both T1-weighted and T2-weighted imaging data are available then the recommendation is to use MIST (Visser et al. 2016), which leverages information from multiple modalities to inform segmentation's. Al-

though MIST can also take diffusion data as an input we have found that if the data is not of a high-enough quality, it could compromise repeatability so we would not recommend this in the first instance as high quality diffusion data which would aid segmentation would probably be acquired in a bespoke manner and differ significantly from what is typically acquired. If mutli-modal data of this kind is not available, FreeSurfer remains the default recommendation.

Tract-wise DTI microstructure metrics

The results from our assessment of the reproducibility of DTI FA measures averaged in white matter ROIs indicated that extracting these metrics from a skeletonised atlas yields more consistent results than extracting them from subject-specific tractography. We would therefore recommend the TBSS tool (Smith et al. 2006). It must be noted that higher levels of reproducibility achieved by this method may be at the expense of subject sensitivity (Heiervang et al. 2006). While results from tractography are more subject-specific, work remains to be done on increasing reproducibility. As advances continue to be made in tractography, this recommendation may change.

rfMRI derived measures

For the denoising of rfMRI data, our recommendation depends on whether training data is available for each scanner involved in the study. If this data is available, FIX (supervised method) (Salimi-Khorshidi et al. 2014) is to be favoured. If this data is not available then ICA-AROMA (unsupervised method) (Pruim, Mennes, van Rooij, Llera, Buitelaar & Beckmann 2015) is to be favoured. Although we have recommended FIX, it's important to note that it is uncommon to have training data available for each scanner involved in a study. Moreover, having training data of a comparable quantity and quality across all scanners involved in a study is even more uncommon. Given these stringent criteria, ICA-AROMA should be considered if one is looking for re-

sults which reflect a more real-life scenario.

Harmonisation Method

Our recommendation for which harmonisation tool to use is data dependent. If there are more than 10 datasets available, then ComBat (Fortin et al. 2017) is the reccomended method. For fewer data sets than this, Neuroharmony (Garcia-Dias et al. 2020) should be used as this is a way of implementing ComBat on indivudal datastets.

6.2 Conclusions

This thesis contributes to better understanding and addressing the lack of consistency across neuroimaging datasets. To facilitate this, we have built a new resource for comprehensively mapping the extent of the problem and objectively evaluating neuroimaging harmonisation approaches. This resource extends previous efforts in a number of ways, by considering more scanners than before (spanning all major vendors), using more modalities than before, having within-scanner variability references and extracting hundreds of imaging features. We based our acquisition protocols on the UK Biobank multi-modal imaging protocol, but adjusted them accordingly for each scanner/site. As implementation and scanner differences only allow a nominal match for acquisition parameters as a whole, we preserved common practice for each scanner in our protocols, allowing a more realistic (and less bespoke) multi-scanner dataset.

Using this resource, we mapped between-scanner variability for a range of multi-modal imaging features and found that this can be up to 5-10 times more than the variability of within-scanner repeats, while bias can be in the order of 10-15 %. We found that the least affected features were derived from T1-weighted and T2-weighted imaging modalities, which were the most re-

producible across scanners by a margin. This was followed by SWI features, then dMRI features and then finally by rfMRI features. We also found that in some cases, the variability of features from a single subject scanned in multiple scanners was comparable to between-subject variability of our study cohort (N=10) scanned in a in a single scanner; and in the worst of cases comparable to the variability of a larger population study (N=1000).

In addition, the scope and size of our travelling-heads dataset enabled us to assess the effects of scanner-induced variance using metrics that have been relatively unexplored. Specifically, we were able to assess consistency between scanners in preserving subject ranking of features. We saw even among features that exhibited the highest degree of between-scanner reproducibility, consistency in subject ranking was not always preserved. This highlights another challenge in harmonising datasets in that features derived from imaging modalities which may have been generally thought to be reproducible and robust to scanner effects still require significant attention.

We used our resource as a testbed to evaluate and objectively compare harmonisation approaches. We demonstrated how adjusting processing steps in pipelines can minimise between-scanner variability in extracted features compared to e.g. within-scanner variability or biological variability (*implicit harmonisation*). We have done this for structural, diffusion and functional imaging modalities. For structural modalities, processing steps have been identified that minimise between-scanner variability for cortical and sub-cortical segmentation of brain regions. For diffusion MRI, processing steps and tools have been identified and denoising methods explored which minimise between-scanner variability of DTI-derived metrics in white matter pathways. For functional MRI, denoising methods have been explored which reduce inter-scanner variability in node amplitudes in ways that can generalise to other vendors.

We have explored and compared the efficacy of two *explicit harmonisation* tools in reducing between-scanner variability. Using within-scanner variability as a baseline reference, we found that harmonisation tools' efficacy varies considerably between feature types, with examples of working very well (e.g. FA values in ROIs) to examples of not causing any major differences (e.g. volumes of subcortical regions). Even so, we found that in cases where harmonisation has reduced variability between scanners it has little to no effect on the preservation of cross subject ranking compared to raw, unharmonised data. So for features where the latter is problematic, explicit harmonisation may not solve the issue. For example, explicit harmonisation would not be a suitable way to remove the variability incurred from scanning 2 different timepoints in 2 different scanners as part of a longitudinal study as this would be susceptible to a change of ranking.

This was in contrast to implicit harmonisation approaches that seem to affect both aspects (reducing between-scanner variability and improving preservation of subject ranking across scanners). A combination of the two is therefore likely to be needed to resolve the harmonisation challenge and our resource can be used to optimise potential solutions.

6.3 Future Perspectives

For current algorithms which claim to harmonise and those yet to come, the developed resource provides objective ways to evaluate them across a number of imaging modalities. The availability of sufficiently rich datasets to train complex models on has led to increased usage of machine learning in MRI (Davatzikos 2019) yet a lack of harmonisation remains a significant obstacle for pooling together datasets that are often acquired across imaging facilities. As the amounts of these datasets continue to increase, harmonisation approaches will inevitably proliferate. Furthermore, there is unlikely to be a one-size-fits

all harmonisation approach. A more likely scenario, as we have seen in Chapter 2, and confirmed with our results in the previous chapter, is that the success of a harmonisation approach will be dependent on the application and type of features.

For example, a multicenter clinical trial would be an ideal candidate for ComBat harmonisation. These studies typically contain a large number of participants (100-1000) and would comfortably satisfy the amount of data sets required for ComBat to operate in an even more optimal way than has been demonstrated in this thesis. On the other hand, in a case-control group comparison, ComBat would be less suitable as it has limited effect on subject-ranking. For example, in conditions characterised by the enlargement or reduction of certain brain regions, (Kang et al. 2020) certain subjects may appear to exhibit a response because of a scanner effect rather than true biological variability.
and as such the multi-modal nature of our resource will help identify the right tool for a given application.

The work presented in this thesis is also an aid in addressing the reproducibility crisis. Various studies have reported that even in cases where the same dataset is being used (Griffanti, Rolinski, Szewczyk-Krolikowski, Menke, Filippini, Zamboni, Jenkinson, Hu & Mackay 2016, Botvinik-Nezer et al. 2020, Schilling et al. 2021), differences in analyses, processing tools and pipelines have yielded inconsistent results. Here, we have shown how processing steps can be objectively compared to reduce between-scanner variability, which is propagated to features and potentially amplified with different magnitude by different processing approaches. Our resource can be used to give recommendations and help achieve consensus on what analyses, processing tools and pipelines should be used. This will increase confidence in comparisons of results made between different research groups and lead to results which reflect

underlying biology of interest and are not confounded by processing tools and pipeline choices.

Another future direction opened up by this work is addressing the challenge of preserving subject ranking across scanners. Harmonisation is typically thought of in terms of reducing between-scanner variance and bias but we have shown that this does not tell the whole-story. ComBat, one of the most popular explicit harmonisation tools, reduces variance and bias for several applications (Fortin et al. 2017, 2018) but we have shown that it has a much reduced effect on restoring cross-subject ranking, which can influence results even in case-control group comparisons. The travelling heads paradigm adopted in our data set and the number of participants included allow future harmonisation algorithms to to be evaluated by their ability to preserve cross-subject ranking and address it accordingly. Interestingly, we found that implicit harmonisation approaches are beneficial in this respect, suggesting that it would be interesting to explore combinations of both groups of methods for optimal harmonisation.

Finally, in acquiring our dataset we have aimed that its scope should represent the kind of variety expected from multi-site studies as we have acquired data from all 3 major vendors and from different generations of scanners from the same vendor. Although we have only acquired data from 10 subjects, this could still be considered a large enough dataset to design and train modern machine learning approaches for certain applications. For instance, when harmonised grey/white matter segmentation is the aim, even cutting-edge deep-learning harmonisation algorithms (Dinsdale et al. 2021) can be trained on 2D data with very high generalisability on 3D datasets. Considering in our dataset slices rather than volumes (which can be further increased further through data augmentation (Chlap et al. 2021) approaches) can enable developments in such a direction.

Bibliography

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E. et al. (2018), 'Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank', *Neuroimage* **166**, 400–424.

Andersson, J. L., Graham, M. S., Zsoldos, E. & Sotiropoulos, S. N. (2016), 'Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion mr images', *Neuroimage* **141**, 556–572.

Andersson, J. L., Jenkinson, M., Smith, S. et al. (2007), 'Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2', *FMRIB Analysis Group of the University of Oxford* **2**(1), e21.

Andersson, J. L., Skare, S. & Ashburner, J. (2003), 'How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging', *Neuroimage* **20**(2), 870–888.

Andersson, J. L. & Sotiropoulos, S. N. (2016), 'An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging', *Neuroimage* **125**, 1063–1078.

Atkinson, D., Hill, D. L., Stoyle, P. N., Summers, P. E. & Keevil, S. F. (1997), 'Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion', *IEEE Transactions on Medical imaging* **16**(6), 903–910.

Badhwar, A., Collin-Verreault, Y., Orban, P., Urchs, S., Chouinard, I., Vogel, J., Potvin, O., Duchesne, S. & Bellec, P. (2020), 'Multivariate consistency of resting-state fmri connectivity maps acquired on a single individual over 2.5 years, 13 sites and 3 vendors', *NeuroImage* **205**, 116210.

Bagherimofidi, S. M., Yang, C. C., Rey-Dios, R., Kanakamedala, M. R. & Fatemi, A. (2019), 'Evaluating the accuracy of geometrical distortion correction of magnetic resonance images for use in intracranial brain tumor radiotherapy', *Reports of Practical Oncology and Radiotherapy* **24**(6), 606–613.

Basser, P. J., Mattiello, J. & LeBihan, D. (1994), 'Mr diffusion tensor spectroscopy and imaging', *Biophysical journal* **66**(1), 259–267.

Bastiani, M., Andersson, J., Cottaar, M., Alfaro-Almagro, F., Fitzgibbon, S. P., Suri, S., Sotiropoulos, S. N. & Jbabdi, S. (2018), 'Eddy qc: Automated quality control for diffusion mri', *International Society for Magnetic Resonance in Medicine, Paris, France* .

Beckmann, C. F. & Smith, S. M. (2004), 'Probabilistic independent component analysis for functional magnetic resonance imaging', *IEEE transactions on medical imaging* **23**(2), 137–152.

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., Linn, K. A., Initiative, A. D. N. et al. (2020), 'Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data', *Neuroimage* **220**, 117129.

Behrens, T. E., Berg, H. J., Jbabdi, S., Rushworth, M. F. & Woolrich, M. W. (2007), 'Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?', *neuroimage* **34**(1), 144–155.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

Borrelli, P., Palma, G., Tedeschi, E., Cocozza, S., Comerci, M., Alfano, B., Haacke, E. M. & Salvatore, M. (2015), 'Improving signal-to-noise ratio in susceptibility weighted imaging: a novel multicomponent non-local approach', *PloS one* **10**(6), e0126835.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A. et al. (2020), 'Variability in the analysis of a single neuroimaging dataset by many teams', *Nature* **582**(7810), 84–88.

Brodmann, K. (1909), *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*, Barth.

Bruno, M. A., Walker, E. A. & Abujudeh, H. H. (2015), 'Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction', *Radiographics* **35**(6), 1668–1676.

Bugge, R., Beyer, M., Nerland, S., Kjonigsen, L., Andreassen, O. & Nordhøy, W. (2017), 'Comparing the t1-weighted sequences mprage and bravo for automated brain segmentation'.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H. et al. (2018), 'The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites', *Developmental cognitive neuroscience* **32**, 43–54.

Castellanos, F. X., Di Martino, A., Craddock, R. C., Mehta, A. D. & Milham, M. P. (2013), 'Clinical applications of the functional connectome', *Neuroimage* **80**, 527–540.

Chalavi, S., Simmons, A., Dijkstra, H., Barker, G. J. & Reinders, A. (2012), 'Quantitative and qualitative assessment of structural magnetic resonance imaging data in a two-center study', *BMC medical imaging* **12**(1), 1–15.

Chang, H. & Fitzpatrick, J. M. (1990), Geometrical image transformation to compensate for mri distortions, *in* 'Medical Imaging IV: Image Processing', Vol. 1233, SPIE, pp. 116–127.

Chen, A. A., Beer, J. C., Tustison, N. J., Cook, P. A., Shinohara, R. T., Shou, H., Initiative, A. D. N. et al. (2020), 'Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data', *bioRxiv* p. 858415.

Chen, C.-C., Wan, Y.-L., Wai, Y.-Y. & Liu, H.-L. (2004), 'Quality assurance of clinical mri scanners using acr mri phantom: preliminary results', *Journal of digital imaging* **17**(4), 279–284.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L. & Haworth, A. (2021), 'A review of medical image data augmentation techniques for deep learning applications', *Journal of Medical Imaging and Radiation Oncology* **65**(5), 545–563.

Chung, A. W., Seunarine, K. K. & Clark, C. A. (2016), 'Noddi reproducibility and variability with magnetic field strength: a comparison between 1.5 t and 3 t', *Human brain mapping* **37**(12), 4550–4565.

Clarke, W. T., Mougin, O., Driver, I. D., Rua, C., Morgan, A. T., Asghar, M., Clare, S., Francis, S., Wise, R. G., Rodgers, C. T. et al. (2020), 'Multi-site harmonization of 7 tesla mri neuroimaging protocols', *NeuroImage* **206**, 116335.

Cox, R. W. (1996), 'Afni: software for analysis and visualization of functional magnetic resonance neuroimages', *Computers and Biomedical research* **29**(3), 162–173.

Cox, R. W. & Hyde, J. S. (1997), 'Software tools for analysis and visualization of fmri data', *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* **10**(4-5), 171–178.

Daducci, A., Canales-Rodríguez, E. J., Zhang, H., Dyrby, T. B., Alexander, D. C. & Thiran, J.-P. (2015), 'Accelerated microstructure imaging via convex optimization (amico) from diffusion mri data', *Neuroimage* **105**, 32–44.

Davatzikos, C. (2019), 'Machine learning in neuroimaging: Progress and challenges', *Neuroimage* **197**, 652.

de Boer, R., Vrooman, H. A., Ikram, M. A., Vernooij, M. W., Breteler, M. M., van der Lugt, A. & Niessen, W. J. (2010), 'Accuracy and reproducibility study of automatic mri brain tissue segmentation methods', *Neuroimage* **51**(3), 1047–1056.

De Groot, M., Vernooij, M. W., Klein, S., Ikram, M. A., Vos, F. M., Smith, S. M., Niessen, W. J. & Andersson, J. L. (2013), 'Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration', *Neuroimage* **76**, 400–411.

Descoteaux, M., Angelino, E., Fitzgibbons, S. & Deriche, R. (2007), 'Regularized, fast, and robust analytical q-ball imaging', *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **58**(3), 497–510.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T. et al. (2006), 'An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest', *Neuroimage* **31**(3), 968–980.

Destrieux, C., Fischl, B., Dale, A. & Halgren, E. (2010), 'Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature', *Neuroimage* **53**(1), 1–15.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M. et al. (2014), 'The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism', *Molecular psychiatry* **19**(6), 659–667.

Dietrich, O., Raya, J. G., Reeder, S. B., Reiser, M. F. & Schoenberg, S. O. (2007*a*), 'Measurement of signal-to-noise ratios in mr images: influence of multichannel coils, parallel imaging, and reconstruction filters', *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **26**(2), 375–385.

Dietrich, O., Raya, J. G., Reeder, S. B., Reiser, M. F. & Schoenberg, S. O. (2007*b*), 'Measurement of signal-to-noise ratios in mr images: influence of multichannel coils, parallel imaging, and reconstruction filters', *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **26**(2), 375–385.

Dinsdale, N. K., Jenkinson, M. & Namburete, A. I. (2021), 'Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal', *NeuroImage* **228**, 117689.

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V. S., Khademi, A., Bartha, R., Bellec, P., Collins, D. L., Descoteaux, M., Hoge, R. et al. (2019), 'The canadian dementia imaging protocol: harmonizing national cohorts', *Journal of Magnetic Resonance Imaging* **49**(2), 456–465.

Duff, E., Zelaya, F., Almagro, F. A., Miller, K. L., Martin, N., Nichols, T. E., Taschler, B., Griffanti, L., Arthofer, C., Wang, C. et al. (2021), 'Reliability of multi-modal mri-derived brain phenotypes for multi-site assessment of neuropsychiatric complications of sars-cov-2 infection', *medRxiv* .

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A. & Gorgolewski, K. J. (2017), 'Mriqc: Advancing the automatic prediction of image quality in mri from unseen sites', *PloS one* **12**(9), e0184661.

Esteban, O., Blair, R. W., Nielson, D. M., Varada, J. C., Marrett, S., Thomas, A. G., Poldrack, R. A. & Gorgolewski, K. J. (2019), 'Crowdsourced mri quality metrics and expert quality annotations for training of humans and machines', *Scientific data* **6**(1), 30.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M. et al. (2019), 'fmriprep: a robust preprocessing pipeline for functional mri', *Nature methods* **16**(1), 111–116.

Farrell, J. A., Landman, B. A., Jones, C. K., Smith, S. A., Prince, J. L., Van Zijl, P. C. & Mori, S. (2007), 'Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging–derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 t', *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **26**(3), 756–767.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S. et al. (2002), 'Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain', *Neuron* **33**(3), 341–355.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D. et al. (2004), 'Automatically parcellating the human cerebral cortex', *Cerebral cortex* **14**(1), 11–22.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J. et al. (2018), 'Harmonization of cortical thickness measurements across scanners and sites', *Neuroimage* **167**, 104–120.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E. et al. (2017), 'Harmonization of multi-site diffusion tensor imaging data', *Neuroimage* **161**, 149–170.

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N. et al. (2016), 'Removing inter-subject technical variability in magnetic resonance imaging studies', *NeuroImage* **132**, 198–212.

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., Gollub, R. L., Lauriello, J., Lim, K. O., Cannon, T. et al. (2008), 'Test–retest and between-site reliability in a multicenter fmri study', *Human brain mapping* **29**(8), 958–972.

Friston, K. (2007), 'A short history of spm', *Statistical parametrical mapping: The analysis of functional brain images* pp. 3–9.

Fujimoto, K., Polimeni, J. R., Van Der Kouwe, A. J., Reuter, M., Kober, T., Benner, T., Fischl, B. & Wald, L. L. (2014), 'Quantitative comparison of cortical surface reconstructions from mp2rage and multi-echo mprage data at 3 and 7 t', *Neuroimage* **90**, 60–73.

Ganzetti, M., Wenderoth, N. & Mantini, D. (2016), 'Intensity inhomogeneity correction of structural mr images: a data-driven approach to define input algorithm parameters', *Frontiers in neuroinformatics* **10**, 10.

Garcia-Dias, R., Scarpazza, C., Baecker, L., Vieira, S., Pinaya, W. H., Corvin, A., Redolfi, A., Nelson, B., Crespo-Facorro, B., McDonald, C. et al. (2020), 'Neuroharmony: A new tool for harmonizing volumetric mri data from unseen scanners', *Neuroimage* **220**.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R. et al. (2013), 'The minimal preprocessing pipelines for the human connectome project', *Neuroimage* **80**, 105–124.

Glocker, B., Robinson, R., Castro, D. C., Dou, Q. & Konukoglu, E. (2019), 'Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects', *arXiv preprint arXiv:1910.04597* .

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O. et al. (2016), 'The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments', *Scientific data* **3**(1), 1–9.

Greve, D. N. & Fischl, B. (2009), 'Accurate and robust brain image alignment using boundary-based registration', *Neuroimage* **48**(1), 63–72.

Griffanti, L., Rolinski, M., Szewczyk-Krolikowski, K., Menke, R. A., Filippini, N., Zamboni, G., Jenkinson, M., Hu, M. T. & Mackay, C. E. (2016), 'Challenges in the reproducibility of clinical studies with resting state fmri: An example in early parkinson's disease', *Neuroimage* **124**, 704–713.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E. et al. (2014), 'Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging', *Neuroimage* **95**, 232–247.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M. et al. (2016), 'Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities', *Neuroimage* **141**, 191–205.

Gronenschild, E. H., Burgmans, S., Smeets, F., Vuurman, E. F., Uylings, H. B. & Jolles, J. (2010), 'A time-saving and facilitating approach for segmentation of anatomically defined cortical regions: Mri volumetry', *Psychiatry Research: Neuroimaging* **181**(3), 211–218.

Gudbjartsson, H. & Patz, S. (1995), 'The rician distribution of noisy mri data', *Magnetic resonance in medicine* **34**(6), 910–914.

Hainline, A. E., Nath, V., Parvathaneni, P., Blaber, J., Rogers, B., Newton, A., Luci, J., Edmonson, H., Kang, H. & Landman, B. A. (2018), Evaluation of inter-site bias and variance in diffusion-weighted mri, *in* 'Medical Imaging 2018: Image Processing', Vol. 10574, SPIE, pp. 266–276.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R. et al. (2006), 'Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer', *Neuroimage* **32**(1), 180–194.

Hawco, C., Viviano, J. D., Chavez, S., Dickie, E. W., Calarco, N., Kochunov, P., Argyelan, M., Turner, J. A., Malhotra, A. K., Buchanan, R. W. et al. (2018), 'A longitudinal human phantom reliability study of multi-center t1-weighted, dti, and resting state fmri data', *Psychiatry Research: Neuroimaging* **282**, 134–142.

Heiervang, E., Behrens, T., Mackay, C. E., Robson, M. D. & Johansen-Berg, H. (2006), 'Between session reproducibility and between subject variability of diffusion mr and tractography measures', *Neuroimage* **33**(3), 867–877.

Hernandez-Fernandez, M., Reguly, I., Jbabdi, S., Giles, M., Smith, S. & Sotiropoulos, S. N. (2019), 'Using gpus to accelerate computational diffusion mri: From microstructure estimation to tractography and connectomes', *Neuroimage* **188**, 598–615.

Hernández, M., Guerrero, G. D., Cecilia, J. M., García, J. M., Inuggi, A., Jbabdi, S., Behrens, T. E. & Sotiropoulos, S. N. (2013), 'Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using gpus', *PloS one* **8**(4), e61892.

Hidalgo-Tobon, S. S. (2010), 'Theory of gradient coil design methods for magnetic resonance imaging', *Concepts in Magnetic Resonance Part A* **36**(4), 223–242.

Iscan, Z., Jin, T. B., Kendrick, A., Szeglin, B., Lu, H., Trivedi, M., Fava, M., McGrath, P. J., Weissman, M., Kurian, B. T. et al. (2015), 'Test–retest reliability of freesurfer measurements within and between sites: Effects of visual approval process', *Human brain mapping* **36**(9), 3472–3485.

Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C. et al. (2008), 'The alzheimer's disease neuroimaging initiative (adni): Mri methods', *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **27**(4), 685–691.

Jbabdi, S., Sotiropoulos, S. N., Savio, A. M., Graña, M. & Behrens, T. E. (2012), 'Model-based analysis of multishell diffusion mr data for tractography: How to get over fitting problems', *Magnetic resonance in medicine* **68**(6), 1846–1855.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. (2012), 'Fsl', *Neuroimage* **62**(2), 782–790.

Ji, J. L., Demšar, J., Fonteneau, C., Tamayo, Z., Pan, L., Kraljič, A., Matkovič, A., Purg, N., Helmer, M., Warrington, S. et al. (2022), 'Qunex–an integrative platform for reproducible neuroimaging analytics', *bioRxiv* .

Johnson, W. E., Li, C. & Rabinovic, A. (2007), 'Adjusting batch effects in microarray expression data using empirical bayes methods', *Biostatistics* **8**(1), 118–127.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van Der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J. et al. (2006), 'Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data', *Neuroimage* **30**(2), 436–443.

Jovicich, J., Minati, L., Marizzoni, M., Marchitelli, R., Sala-Llonch, R., Bartrés-Faz, D., Arnold, J., Benninghoff, J., Fiedler, U., Roccatagliata, L. et al. (2016), 'Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: a multicentric resting-state fmri study', *Neuroimage* **124**, 442–454.

Kang, K., Han, J., Lee, S.-W., Jeong, S. Y., Lim, Y.-H., Lee, J.-M. & Yoon, U. (2020), 'Abnormal cortical thickening and thinning in idiopathic normal-pressure hydrocephalus', *Scientific Reports* **10**(1), 21213.

Karayumak, S. C., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M. & Rathi, Y. (2019), 'Retrospective harmonization of multi-site diffusion mri data acquired with different acquisition parameters', *Neuroimage* **184**, 180–200.

Karayumak, S. C., Kubicki, M. & Rathi, Y. (2019), 'Multi-site diffusion mri harmonization in the presence of gross pathology: How far can we push?'.

Keenan, K. E., Stupic, K. F., Boss, M. A., Russek, S. E., Chenevert, T. L., Prasad, P. V., Reddick, W. E., Zheng, J., Hu, P., Jackson, E. F. et al. (2016), 'Comparison of t1 measurement using ismrm/nist system phantom'.

Keil, B., Blau, J. N., Biber, S., Hoecht, P., Tountcheva, V., Setsompop, K., Triantafyllou, C. & Wald, L. L. (2013), 'A 64-channel 3t array coil for accelerated brain mri', *Magnetic resonance in medicine* **70**(1), 248–258.

Knussmann, G. N., Anderson, J. S., Prigge, M. B., Dean III, D. C., Lange, N., Bigler, E. D., Alexander, A. L., Lainhart, J. E., Zielinski, B. A. & King, J. B. (2022), 'Test-retest reliability of freesurfer-derived volume, area and cortical thickness from mprage and mp2rage brain mri images', *Neuroimage: Reports* **2**(2), 100086.

Koike, S., Tanaka, S. C., Okada, T., Aso, T., Yamashita, A., Yamashita, O., Asano, M., Maikusa, N., Morita, K., Okada, N. et al. (2021), 'Brain/minds beyond human brain mri project: a protocol for multi-level harmonization across brain disorders throughout the lifespan', *NeuroImage: Clinical* **30**, 102600.

Krüger, G. & Glover, G. H. (2001), 'Physiological noise in oxygenation-sensitive magnetic resonance imaging', *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **46**(4), 631–637.

Kurokawa, R., Kamiya, K., Koike, S., Nakaya, M., Uematsu, A., Tanaka, S. C., Kamagata, K., Okada, N., Morita, K., Kasai, K. et al. (2021), 'Cross-scanner reproducibility and harmonization of a diffusion mri structural brain network: A traveling subject study of multi-b acquisition', *NeuroImage* **245**, 118675.

Laun, F. B., Huff, S. & Stieltjes, B. (2009), 'On the effects of dephasing due to local gradients in diffusion tensor imaging experiments: relevance for diffusion tensor imaging fiber phantoms', *Magnetic resonance imaging* **27**(4), 541–548.

Lauterbur, P. C. (1973), 'Image formation by induced local interactions: examples employing nuclear magnetic resonance', *nature* **242**(5394), 190–191.

Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. (2016), 'The first step for neuroimaging data analysis: Dicom to nifti conversion', *Journal of neuroscience methods* **264**, 47–56.

Liney, G., Owen, S., Beaumont, A., Lazar, V., Manton, D. & Beavis, A. (2013), 'Commissioning of a new wide-bore mri scanner for radiotherapy planning of head and neck cancer', *The British Journal of Radiology* **86**(1027), 20130150.

Magnotta, V. A. & Friedman, L. (2006), 'Measurement of signal-to-noise and contrast-to-noise in the fbirn multicenter imaging study', *Journal of digital imaging* **19**(2), 140–147.

Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S. C. et al. (2021), 'Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics', *Human brain mapping* **42**(16), 5278–5287.

Makris, N., Goldstein, J. M., Kennedy, D., Hodge, S. M., Caviness, V. S., Faraone, S. V., Tsuang, M. T. & Seidman, L. J. (2006), 'Decreased volume of left and total anterior insular lobule in schizophrenia', *Schizophrenia research* **83**(2-3), 155–171.

Mansfield, P. & Maudsley, A. A. (1977), 'Medical imaging by nmr', *The British journal of radiology* **50**(591), 188–194.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. et al. (2016), 'Multimodal population brain imaging in the uk biobank prospective epidemiological study', *Nature neuroscience* **19**(11), 1523–1536.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C. E., Morey, R. A., Flashman, L. A. et al. (2016), 'Inter-site and inter-scanner diffusion mri data harmonization', *NeuroImage* **135**, 311–323.

Molloy, E. K., Meyerand, M. E. & Birn, R. M. (2014), 'The influence of spatial resolution and smoothing on the detectability of resting-state and task fmri', *Neuroimage* **86**, 221–230.

Mori, S., Wakana, S., Van Zijl, P. C. & Nagae-Poetscher, L. (2005), *MRI atlas of human white matter*, Elsevier.

Mortamet, B., Bernstein, M. A., Jack Jr, C. R., Gunter, J. L., Ward, C., Britson, P. J., Meuli, R., Thiran, J.-P. & Krueger, G. (2009), 'Automatic quality assessment in structural brain magnetic resonance imaging', *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **62**(2), 365–372.

Moyer, D., Ver Steeg, G., Tax, C. M. & Thompson, P. M. (2020), 'Scanner invariant representations for diffusion mri harmonization', *Magnetic resonance in medicine* **84**(4), 2174–2189.

Nyúl, L. G. & Udupa, J. K. (1999), 'On standardizing the mr image intensity scale', *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **42**(6), 1072–1081.

Orlhac, F., Nioche, C., Klyuzhin, I., Rahmim, A. & Buvat, I. (2021), 'Radiomics in pet imaging: a practical guide for newcomers', *PET clinics* **16**(4), 597–612.

Oztek, M. A., Brunnquell, C. L., Hoff, M. N., Boulter, D. J., Mossa-Basha, M., Beauchamp, L. H., Haynor, D. L. & Nguyen, X. V. (2020), 'Practical considerations for radiologists in implementing a patient-friendly mri experience', *Topics in Magnetic Resonance Imaging* **29**(4), 181–186.

Pagani, E., Hirsch, J. G., Pouwels, P. J., Horsfield, M. A., Perego, E., Gass, A., Roosendaal, S. D., Barkhof, F., Agosta, F., Rovaris, M. et al. (2010), 'Intercenter differences in diffusion tensor mri acquisition', *Journal of Magnetic Resonance Imaging* **31**(6), 1458–1468.

Palacios, E. M., Martin, A. J., Boss, M. A., Ezekiel, F., Chang, Y. S., Yuh, E. L., Vassar, M. J., Schnyer, D. M., MacDonald, C. L., Crawford, K. L. et al. (2017), 'Toward precision and reproducibility of diffusion tensor imaging: a multicenter diffusion phantom and traveling volunteer study', *American Journal of Neuroradiology* **38**(3), 537–545.

Patenaude, B., Smith, S. M., Kennedy, D. N. & Jenkinson, M. (2011), 'A bayesian model of shape and appearance for subcortical brain segmentation', *Neuroimage* **56**(3), 907–922.

Pohl, K. M., Sullivan, E. V., Rohlfing, T., Chu, W., Kwon, D., Nichols, B. N., Zhang, Y., Brown, S. A., Tapert, S. F., Cummins, K. et al. (2016), 'Harmonizing dti measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the ncanda study', *Neuroimage* **130**, 194–213.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y. et al. (2020), 'Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan', *NeuroImage* **208**, 116450.

Pruim, R. H., Mennes, M., Buitelaar, J. K. & Beckmann, C. F. (2015), 'Evaluation of ica-aroma and alternative strategies for motion artifact removal in resting state fmri', *Neuroimage* **112**, 278–287.

Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K. & Beckmann, C. F. (2015), 'Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data', *Neuroimage* **112**, 267–277.

Rao, A., Monteiro, J. M., Mourao-Miranda, J., Initiative, A. D. et al. (2017), 'Predictive modelling using neuroimaging data in the presence of confounds', *Neuroimage* **150**, 23–49.

Reig, S., Sánchez-González, J., Arango, C., Castro, J., González-Pinto, A., Ortuno, F., Crespo-Facorro, B., Bargallo, N. & Desco, M. (2009), 'Assessment of the increase in variability when combining volumetric data from different scanners', *Human brain mapping* **30**(2), 355–368.

Rorden, C., Morgan, P. S., Parekh, P. & Bhalerao, G. (2012), 'Calculating totalreadouttime for philips mri', *Neuroimage* **62**(2), 782–790.

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L. & Smith, S. M. (2014), 'Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers', *Neuroimage* **90**, 449–468.

Schilling, K. G., Rheault, F., Petit, L., Hansen, C. B., Nath, V., Yeh, F.-C., Girard, G., Barakovic, M., Rafael-Patino, J., Yu, T. et al. (2021), 'Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset?', *NeuroImage* **243**, 118502.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M. et al. (2014), 'Statistical normalization techniques for magnetic resonance imaging', *NeuroImage: Clinical* **6**, 9–19.

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

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M. et al. (2006), 'Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data', *Neuroimage* **31**(4), 1487–1505.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E. et al. (2004), 'Advances in functional and structural mr image analysis and implementation as fsl', *Neuroimage* **23**, S208–S219.

Smith, S. M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P. M., Federico, A. & De Stefano, N. (2002), 'Accurate, robust, and automated longitudinal and cross-sectional brain change analysis', *Neuroimage* **17**(1), 479–489.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. et al. (2015), 'Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age', *PLoS medicine* **12**(3), e1001779.

Takao, H., Hayashi, N. & Ohtomo, K. (2011), 'Effect of scanner in longitudinal studies of brain volume changes', *Journal of Magnetic Resonance Imaging* **34**(2), 438–444.

Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A. et al. (2021), 'A multi-site, multi-disorder resting-state magnetic resonance image database', *Scientific data* **8**(1), 1–15.

Tax, C. M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., Evans, C. J., St-Jean, S., Leemans, A., Koppers, S., Merhof, D. et al. (2019), 'Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms', *NeuroImage* **195**, 285–299.

Torralba, A. & Efros, A. A. (2011), Unbiased look at dataset bias, *in* 'CVPR 2011', IEEE, pp. 1521–1528.

Treit, S., Stolz, E., Rickard, J. N., McCreary, C. R., Bagshawe, M., Frayne, R., Lebel, C., Emery, D. & Beaulieu, C. (2022), 'Lifespan volume trajectories from non–harmonized t1–weighted mri do not differ after site correction based on traveling human phantoms', *Frontiers in Neurology* **13**.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H. et al. (2013), 'The wu-minn human connectome project: an overview', *Neuroimage* **80**, 62–79.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W. et al. (2012), 'The human connectome project: a data acquisition perspective', *Neuroimage* **62**(4), 2222–2231.

Van Horn, J. D. & Toga, A. W. (2009), 'Multi-site neuroimaging trials', *Current opinion in neurology* **22**(4), 370.

Venkatraman, V. K., Gonzalez, C. E., Landman, B., Goh, J., Reiter, D. A., An, Y. & Resnick, S. M. (2015), 'Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths', *Neuroimage* **119**, 406–416.

Veraart, J., Novikov, D. S., Christiaens, D., Ades-Aron, B., Sijbers, J. & Fieremans, E. (2016), 'Denoising of diffusion mri using random matrix theory', *Neuroimage* **142**, 394–406.

Visser, E., Keuken, M. C., Douaud, G., Gaura, V., Bachoud-Levi, A.-C., Remy, P., Forstmann, B. U. & Jenkinson, M. (2016), 'Automatic segmentation of the striatum and globus pallidus using mist: Multimodal image segmentation tool', *NeuroImage* **125**, 479–497.

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P. & Koepp, M. J. (2010), 'Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 t scanners', *Neuroimage* **51**(4), 1384–1394.

Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A. D. N. et al. (2021), 'Detect and correct bias in multi-site neuroimaging datasets', *Medical Image Analysis* **67**, 101879.

Wakana, S., Caprihan, A., Panzenboeck, M. M., Fallon, J. H., Perry, M., Gollub, R. L., Hua, K., Zhang, J., Jiang, H., Dubey, P. et al. (2007), 'Reproducibility of quantitative tractography methods applied to cerebral white matter', *Neuroimage* **36**(3), 630–644.

Warrington, S., Bryant, K. L., Khrapitchev, A. A., Sallet, J., Charquero-Ballester, M., Douaud, G., Jbabdi, S., Mars, R. B. & Sotiropoulos, S. N. (2020), 'Xtract-standardised protocols for automated tractography in the human and macaque brain', *Neuroimage* **217**, 116923.

Wrobel, J., Martin, M., Bakshi, R., Calabresi, P. A., Elliot, M., Roalf, D., Gur, R. C., Gur, R. E., Henry, R. G., Nair, G. et al. (2020), 'Intensity warping for multisite mri harmonization', *NeuroImage* **223**, 117242.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N. et al. (2019), 'Harmonization of resting-state functional mri data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias', *PLoS biology* **17**(4), e3000042.

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T. & Sheline, Y. I. (2018), 'Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data', *Human brain mapping* **39**(11), 4213–4227.

Zaitsev, M., Maclaren, J. & Herbst, M. (2015), 'Motion artifacts in mri: A complex problem with many partial solutions', *Journal of Magnetic Resonance Imaging* **42**(4), 887–901.

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. (2012), 'Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain', *Neuroimage* **61**(4), 1000–1016.

Zhang, Y., Brady, M. & Smith, S. (2001), 'Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm', *IEEE transactions on medical imaging* **20**(1), 45–57.

Zhong, J., Wang, Y., Li, J., Xue, X., Liu, S., Wang, M., Gao, X., Wang, Q., Yang, J. & Li, X. (2020), 'Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development', *Biomedical engineering online* **19**(1), 1–18.

Zhou, F.-L., Li, Z., Gough, J. E., Cristinacce, P. L. H. & Parker, G. J. (2018), 'Axon mimicking hydrophilic hollow polycaprolactone microfibres for diffusion magnetic resonance imaging', *Materials & design* **137**, 394–403.

Zhu, T., Hu, R., Qiu, X., Taylor, M., Tso, Y., Yiannoutsos, C., Navia, B., Mori, S., Ekholm, S., Schifitto, G. et al. (2011), 'Quantification of accuracy and precision of multi-center dti measurements: a diffusion phantom and human brain study', *Neuroimage* **56**(3), 1398–1411.