

A thesis submitted for the degree of Doctor of Philosophy

Brain imaging in UK Biobank

Fidel Alfaro-Almagro

Pembroke College



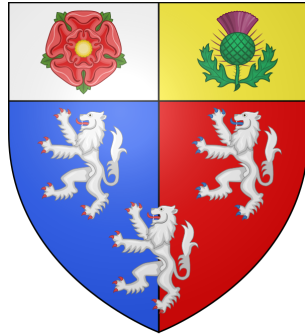
FMRIB Analysis Group

Wellcome Centre for Integrative Neuroimaging (WIN)

Nuffield Department of Clinical Neurosciences

University of Oxford

Trinity 2020



Pembroke College



Wellcome Centre for Integrative Neuroimaging (WIN - FMRIB)



UK Biobank

ABSTRACT

Brain imaging in UK Biobank

UK Biobank is a large-scale prospective epidemiological study with all data accessible to researchers worldwide. It is currently in the process of bringing back 100,000 of the original participants for brain, heart and body MRI, carotid ultrasound and low-dose bone/fat x-ray. The brain imaging component covers six modalities (T1, T2 FLAIR, susceptibility-weighted MRI, Resting fMRI, Task fMRI and Diffusion MRI). Raw and processed data from the first $\sim 42,000$ imaged subjects has recently been released for general research access. To help convert this data into useful summary information, we have developed an automated processing and QC (Quality Control) pipeline that is available for use by other researchers. In this work, we describe the tools and research we have developed in order to generate the data in a way that is readily usable for researchers worldwide. The first half of this thesis will give a brief overview of UK Biobank brain imaging and the acquisition protocol, and then describe the processing pipeline and QC pipeline in detail. We also describe several quantitative investigations carried out as part of the development of both the imaging protocol and the processing pipeline. The second part of this thesis will show the research we performed to come up with a set of recommendations to deal with confounds in brain imaging studies using UK Biobank. UK Biobank is a powerful resource for studying associations between imaging and non-imaging measures such as lifestyle factors and health outcomes, in part because of the large subject numbers. However, the resulting high statistical power also raises the sensitivity to confounding effects, which therefore have to be carefully considered. Here we describe a set of possible confounds (including non-linear effects and interactions) that researchers may wish to consider for their studies using such data. We include descriptions of how we can estimate the confounds, and study the extent to which each of these confounds affects the data, and the spurious correlations that may arise if they are not controlled.

ACKNOWLEDGEMENTS

A mi familia y amigos de España, sin los que nada de esto hubiera sido posible.

To my doctoral advisors, Prof. Stephen M. Smith, Prof. Thomas E. Nichols, and Prof. Mark Jenkinson, for being at the same time extraordinarily knowledgeable, and eternally patient in their mentoring. If I ever advise a (poor) PhD student, I hope I will be at least half as kind and inspiring as they were.

I am very much indebted to UK Biobank. First and foremost, to the thousands of volunteers, for selflessly making this incredible resource possible. Second, to the project itself and its promoters, for their foresight in putting together such an incredible opportunity for the whole world to use. I am absolutely thankful for having been funded eight years by this project and for the trust I have received. I hope this piece of work shows my immense gratitude. I would also like to acknowledge the UK Medical Research Council, and the Wellcome Trust for funding of UK biobank, and the Wellcome Trust UK Strategic Award [098369/Z/12/Z], and other core funding from the Wellcome Trust (203139/Z/16/Z) for funding at FMRI.

Doing a thesis is a wild adventure, and I could not have had better fellow travellers in this trip than the ones in The Office: Ludovica Griffanti, Jelena Božek, Paul McCarthy, Sean Fitzgibbon, Matthew Webster, Taylor Hanayik, Ged Ridgway, and Tom Whyntie. Here's to the endless support, friendship, and biscuits.

Working and studying at the University of Oxford has offered me the incredible opportunity to learn from some of the most brilliant minds in my field. I cannot stress strongly enough how much I have learnt from Professors Karla Miller, Saad Jbabdi, Gwenaëlle Douaud, Jesper Andersson, Clare Mackay, Christian Beckmann, Heidi Johansen-Berg, Tim Behrens and Irene Tracey. There is a big list of out-of-Oxford collaborators whom I would also like to thank: Daniel Alexander, Martina Callaghan, Alessandro Daducci, Iulius Dragonu, Mark Effingham, Lloyd Elliott, Oscar Esteban, Bruce Fischl, Steve Garratt, Chris Gorgolewski, Douglas Greve, Takuya Hayashi, Jonathan Marchini, Paul Matthews, John Miller, Niels Oesingmann, Russ Poldrack, Jonathan Price, Chris Rorden, Gholamreza Salimi-Khorshidi, Alan Young, and Hui Zhang. Additionally, I would like to thank Susan Field, for her invaluable work and for being my (caffeine) dealer for seven years.

I have been really privileged to share these years in Oxford with an incredible group of friends and colleagues, listed here in no particular order: Moisés Hernández, Anderson Winkler, Adriana Roca, Kyoko Morimoto, Liz Lee, Soroosh Afyouni, Zobair Arya,

Matteo Bastiani, Janine Bijsterbosch, Lisa Bruckert, Marina Charquero, Mark Chiew, Alexa Constantinescu, Michiel Cottaar, Rezvan Farahibozorg, Geoffrey Ferrari, Nicola Filippini, David Flitney, Francesca Galassi, Jonathan Hadida, Liv Harrison, Samuel Harrison, Ionela Jula, Slava Karolis, Joe Yazhuo Kong, Silvia Messina, Duncan Mortimer, Hamed Nili, Hanna Nowicka, Eugene Duff, Oiwi Parker Jones, Piergiorgio Salvan, Enrico Schulz, Stamatios Sotiropoulos, Vaanathi Uma Sundaresan, Emmanuel Vallee, Johanna Vannesjö, Diego Vidaurre, Eelke Visser, Natalie Voets, and the football crew. Special mention to the many hours spent in The White Hart!

My time in the UK would not have been so good if it were not for the people in the Society of Spanish Researchers in the UK (SRUK): Ruben and Olivia (esa directiva, oé), Alberto, Carmen, Sergio, Pablo, Ana, Sergio (el otro), Carlos, Helena, Cristina, Edu Oliver, Javier, Estrella, Lorenzo, Jose P., and Laura.

Looking back in time, we can always see how many people influenced the path we finally took. For example, the first time I realised scientific research was my real calling was during my time at CERN. I spent some of the best days of my life with Valentina, Juan Maikel, Elena Mikaela, Pablito, Fiorella, Alberto, Piero, Maria Elena, Jonatan, Isabel, Robin, Jorge, Raul MejiMaikel, Weina, Consuelo, Estebanito, Luis, Luisfer, Josemi, and Don Diego. I hope we can get together again soon.

I also fondly remember the friends I made when I decided to take the neuroimaging path and started working in Madrid, at the Medical Imaging Lab (LIM): Yasser, Susanna, Javi, Maria, Chema, Cris, Eu, Paula, Rigo, Sisni, Esther, Trajana, Lorena, Joost, Juan, Pedro, Marta, Kenia, Santi, Diego, Pajarito, Asier, Bea, Claudia, Judit, Ivan, Juanjo (chico), Ana, Natalia, Vero, y Laura. They are my kindred people and taught me all I know about how science is much more than spending endless hours in the lab (better to spend them in La Santina, of course). I would also like to thank the people out of this list because I have learnt from them what science should never be.

Andy suffered many of my ups and downs through most of this process; for her unconditional support, I will be eternally grateful.

There are many people out of these two pages who made this thesis possible by making my life better. It is simply not possible to include everyone here, so I apologise if your name was (unfairly) left out. I owe you a beer.

Finally, to science. For being my light.

Úbeda, September 7, 2020.

Contents

Preface	19
1 Introduction - The UK Biobank	23
1.1 General description of UK Biobank	23
1.2 Imaging studies	25
1.3 Brain imaging study	27
1.3.1 Acquisition overview	28
1.3.2 Acquired modalities overview	30
1.3.3 Protocol evolution	32
1.3.4 Data release history	34
2 Image Processing Pipeline for UK Biobank	37
2.1 Introduction: Processing brain imaging data	37
2.1.1 Imaging-Derived Phenotypes (IDPs)	38
2.2 Automated processing pipelines	41
2.2.1 Initial preprocessing steps	41
2.2.1.1 Data conversion: DICOM to NIFTI	44
2.2.1.2 Data organization	44
2.2.1.3 Image anonymisation	47
2.2.1.4 Gradient distortion correction	48
2.2.2 T1 pipeline	48
2.2.3 T2 FLAIR pipeline	54
2.2.4 swMRI pipeline	55
2.2.5 Fieldmap generation pipeline	61
2.2.6 Fieldmap magnitude generation	63
2.2.7 dMRI pipeline	64
2.2.7.1 Diffusion-Tensor-Image fitting and NODDI	64
2.2.7.2 WM tract skeleton analysis	67
2.2.7.3 Probabilistic-tractography-based analysis	72
2.2.8 Resting fMRI pipeline	73
2.2.9 Task fMRI pipeline	79
2.3 Results: UK Biobank as a resource	82
2.4 Discussion	83
3 Quality Control for UK Biobank	87
3.1 Introduction: The need for Quality Control	87
3.2 Usability rules	89
3.3 QC framework	90

3.3.1	AutoQC for T1	94
3.3.1.1	AutoQC categories	94
3.3.1.2	AutoQC features	98
3.3.1.3	AutoQC tool description	103
3.3.1.4	Performance of autoQC	105
3.3.1.5	Size of the training dataset for QC	106
3.3.2	Outlier detection for QC	107
3.3.3	Structural motion metric	108
3.3.3.1	Dependent variable: Manual measure of motion	108
3.3.3.2	Independent variables	108
3.3.3.3	Linear regression - Fitting and validation	109
3.3.3.4	Further validation	111
3.4	QC on FreeSurfer	113
3.4.1	Description	113
3.4.2	FreeSurfer QC Results	114
3.4.2.1	Removing temporal pole data from DKT atlas	114
3.4.3	T1 vs T1 + T2 FLAIR	115
3.5	Results	120
3.6	Discussion	122
4	Confound modelling	125
4.1	Introduction: Dealing with confounds	125
4.2	Methods	126
4.2.1	Linear model unconfounding & Variance measures	126
4.2.2	Data description	127
4.2.3	Overview of confound types	128
4.2.4	Description of basic confound variables (80 variables)	128
4.2.4.1	Subject-specific confounds (4 variables)	128
4.2.4.2	Scanner / acquisition / processing parameters (20 variables)	130
4.2.4.3	Head Motion (48 variables)	131
4.2.4.4	Table position (8 variables)	133
4.2.5	Processing of confound regressors (602 variables)	134
4.2.5.1	Basic confounds (80 variables, expanded to 240)	135
4.2.5.2	Non-linear transformations (158 variables)	137
4.2.5.3	Crossed terms (84 variables)	138
4.2.5.4	Date/time-related confounds (120 variables)	138
4.2.6	Null evaluations (Random Gaussian confounds)	139
4.2.7	Correlations between IDPs and non-IDPs variables	139
4.2.8	Additive vs non-additive confound effects	140
4.2.9	Comparison with “simple” confounds	141
4.3	Results	143
4.3.1	Non-linear confounds selection	143
4.3.2	Crossed terms selection	145
4.3.3	Variance of IDPs explained by confound groups	147
4.3.4	Effect of unconfounding on IDP-nIDP correlations	149
4.3.5	Importance of non-additive terms	150
4.3.6	Effects of unconfounding with different sets of confounds	150
4.3.7	Date and time confounds	153

4.4	Discussion	153
4.4.1	General considerations	153
4.4.2	Confounds and non-IDP variables	159
4.4.3	Number of confounds	160
4.4.4	Berkson's paradox	161
4.4.5	Sex-split unconfounding	161
4.4.6	Availability of the confounds	162
4.5	Conclusions	163
5	Discussion and future work	165
5.1	Summary	165
5.2	Impact on neuroscience research	166
5.3	Future work: Pipeline	168
5.3.1	Strategic changes in the pipeline	168
5.3.2	General changes in the pipeline processing	169
5.3.3	Initial / final Processing	170
5.3.4	T1	170
5.3.5	T2 FLAIR	170
5.3.6	swMRI	171
5.3.7	FreeSurfer	171
5.3.8	Fieldmap	171
5.3.9	dMRI	172
5.3.10	fMRI	172
5.4	Future work: QC	172
5.5	Future work: Confounds	173
Appendix		175
A.1	Motion and SNR in the imaging modalities	175
A.2	Evaluation of change in shimming	176
A.3	Acquisition and protocol features	179
A.4	Raw confound processing algorithms	183
A.5	Non-additive terms (NATs) processing algorithm	185
A.6	Temporal components	186
A.7	Violin plots	197
A.8	Bland-Altman plots	207
A.9	%VE of IDPs	212
A.10	Manhattan plots	220
A.11	BA plots of unconfounding sets	223
A.12	Eigenspectrum for the Principal Components of the full confound set	224
Bibliography		225

List of Figures

2.1	Flowcharts legend	42
2.2	General flowchart for the pipeline.	43
2.3	Example of defacing	48
2.4	Gradient Distortion Correction examples for different head positions	49
2.5	Flowchart for the T1 processing pipeline.	50
2.6	Modified dilated brain mask to improve non-linear registration	52
2.7	Flowchart for the defacing of the T1.	53
2.8	Flowchart for the T2 FLAIR processing pipeline.	56
2.9	Flowchart for the defacing of the T2 FLAIR	57
2.10	Comparison of different bias field estimation methods	58
2.11	Flowchart for the swMRI processing pipeline.	59
2.12	Flowchart for the fieldmap generation pipeline.	62
2.13	Flowchart for averaging of B0s	65
2.14	Flowchart for the dMRI processing pipeline.	66
2.15	Comparison of 14 different alignment methods of FA to MNI space	68
2.16	Comparison of alignment for tract 1 to 8	69
2.17	Comparison of alignment for tract 9 to 16	70
2.18	Comparison of alignment for tract 17 to 24	71
2.19	Comparison of alignment for tract 24 to 27	72
2.21	Tract similarity evaluation	74
2.22	Flowchart for the fMRI processing pipeline.	75
2.23	Flowchart for the FEAT part of the pipeline.	76
2.24	Group-level activation maps for faces-shapes	81
2.25	Group ME-Z scatter-plots for the three main contrasts	81
2.26	Single-subject activation summary results	82
2.27	Associations between IDPs and non-IDPs	84
2.28	White Matter Hyperintensities detection and importance	85
3.1	Examples of head motion severity for repeated T1w images	91
3.2	Quality effects of reacquiring T1w images	92
3.3	Boxplots with the Z scores for the 888 not-rfMRI IDPs	97
3.4	Examples of problematic (“unusable”) datasets	98
3.5	QC feature: Volume of different tissues on the edges of brain mask	101
3.6	QC feature: Volume of edges with Canny filter	101
3.7	QC feature: Brain mask boundary	103
3.8	QC feature: Distance between borders	104
3.9	Increase in False Negatives when training size decreases	107
3.10	Histogram for the structure motion for unusable subjects	111
3.11	Histogram correlations of all non-connectivity IDPs with the new metric. .	112
3.12	Qoala-T results on all FreeSurfer Datasets	114
3.13	T1 vs T1 + T2 FLAIR - FS volumes of subcortical subsegmentations . . .	116

3.14	T1 vs T1 + T2 FLAIR - Volumes, areas and thickness with 2 atlases - 1	117
3.15	T1 vs T1 + T2 FLAIR - Volumes, areas and thickness with 2 atlases - 2	118
3.16	T1 vs T1 + T2 FLAIR - Additional IDPs	119
4.1	Matrix of variance of each group of confounds explained by each other group	129
4.2	Relationship between Table Position and the Eddy Currents	134
4.3	Non-linear confounds selection	144
4.4	Crossed terms selection selection	146
4.5	Violin plots with % UVE of IDPs by each group of confounds	148
4.6	% VE and % UVE of IDPs by Site confound	149
4.7	BA plots for the correlations of IDPs with Body and Cognitive nonIDPs	151
4.8	Effect of modelling non-additive terms	152
4.9	Violin plot % VE of all IDPs by different sets of confounds	154
4.10	Manhattan plot of the correlations between IDPs and non-IDPs	155
4.11	BA plots to compare two different unconfounding settings	156
4.12	Time and Date PCA results	157
A.1	Motion estimation for the temporal modalities	175
A.2	Inverted SNR and inverted tSNR for different modalities	176
A.3	Shimming change tests	178

List of Tables

1.1	UK Biobank brain MRI protocol	29
2.1	Comparison of normalisation methods for WM intensity	60
2.2	UK Biobank usable modalities	82
3.1	Hierarchy of modalities for unusability	93
3.2	Classification of problems / imperfections for T1	95
3.3	Classification of problems / imperfections per modality in UK Biobank	96
3.4	Cost matrix for MultiCost classifier	105
3.5	Automated QC tool performance for T1 images	106
3.6	Numeric value for QC category	108
3.7	Number of subjects per QC category	109
3.8	MSE for the tested regression methods	110
3.9	Learned weight per feature	110
3.10	Variables most strongly correlated with the new metric	112
3.11	QC results per site	120
3.12	QC percentages per site	120
3.13	QC results per processing batch	121
3.14	QC percentages per processing batch	121
3.15	QC results for FreeSurfer	122
3.16	Discrepancy in QC categories between visits	124
4.1	Size of confound groups before and after processing	135
A.1	Number of subjects per Site	179
A.2	Number of subjects per Processing batch	179
A.3	Number of subjects per CMRR version	179
A.4	Number of subjects per Service Pack	179
A.5	Number of subjects per Protocol version	179
A.6	Number of subjects per Ramp-down event	180
A.7	Number of subjects per Head coil change event	180
A.8	Number of subjects per Cold head change event	180
A.9	Number of subjects per Miscellaneous event	180
A.10	Number of subjects per Flipped SWI	181
A.11	Number of subjects per FreeSurfer processing mode	181
A.12	Number of subjects per Eddy version	181
A.13	Number of subjects per Echo time for rfMRI	181
A.14	Number of subjects per Echo time for tfMRI	181
A.15	Number of subjects per Intensity scaling factor in T1	181
A.16	Number of subjects per Intensity scaling factor in T2 FLAIR	182
A.17	Number of subjects per Intensity scaling factor in swMRI	182
A.18	Number of subjects per Intensity scaling factor in dMRI	182

A.19 Number of subjects per Intensity scaling factor in rfMRI	182
A.20 Number of subjects per Intensity scaling factor in tfMRI	182

List of algorithms

1	swMRI management logic	45
2	T1w images management logic	46
3	fMRI management logic	46
4	dMRI management logic	47
1	Quantitative confounds processing	183
2	Qualitative confounds processing	184
3	NAT processing	185

Acronyms

ADNI	Alzheimer's Disease Neuroimaging Initiative
AP	Anterior/Posterior
AMICO	Accelerated Microstructure Imaging via Convex Optimization
BIDS	Brain Imaging Data Structure
BBR	Boundary-Based Registration.
BA	Bland-Altman
BET	Brain Extraction Tool
BIANCA	Brain Intensity AbNormality Classification Algorithm
CMRR	Center for Magnetic Resonance Research (U Minnesota)
CSF	Cerebrospinal Fluid
DICOM	Digital Imaging and Communications in Medicine
dMRI	Diffusion MRI
DTI	Diffusion Tensor Imaging
ECG	Electrocardiogram
EPI	Echo Planar Imaging
FA	Fractional Anisotropy
FAST	FMRIB's Automated Segmentation Tool
FEAT	fMRI Expert Analysis Tool
FIX	FMRIB's ICA-based Xnoiseifier
FLIRT	FMRIB's Linear Image Registration Tool
fMRI	Functional MRI
FMRIB	Functional Magnetic Resonance Imaging of the Brain
FNIRT	FMRIB's Non-linear Image Registration Tool
FNR	False Negative Rate
FOV	Field-of-View
FPR	False Positive Rate
FSL	FMRIB Software Library
FS	FreeSurfer
GDC	Gradient Distortion Correction
GM	Grey Matter
GPU	Graphics Processing Unit
HCP	Human Connectome Project
ICA	Independent Component Analysis
IDPs	Imaging-derived phenotypes

IQR	Inter-quartile range
MAD	Mean Absolute Deviation
MNI	Montreal Neurological Institute
MRI	Magnetic Resonance Imaging
nIDPs	Non-imaging-derived phenotypes
NHS	National Health System
NIFTI	Neuroimaging Informatics Technology Initiative
NODDI	Neurite orientation dispersion and density imaging
non-IDPs	nIDPs
PA	Posterior/Anterior
PCs	Principal Components
PCA	Principal Component Analysis
QA	Quality Assurance
QC	Quality Control
QN	Quantile Normalization
QUAD	Quality Assessment for DMRI
rfMRI	Resting state fMRI
RF	Radio Frequency
ROI	Region Of Interest
RSN	Resting-State Network
SBREF	Single-Band Reference
SIENA	Structural Image Evaluation, using Normalisation, of Atrophy
SNR	Signal-to-Noise ratio
SWI	swMRI
swMRI	Susceptibility Weighted MRI
T1w	T1 Weighted imaging
T2-FLAIR	T2-weighted-Fluid-Attenuated Inversion Recovery
TBSS	Tract-based spatial statistics
tfMRI	Task fMRI
VE	Variance Explained
UVE	Unique Variance Explained
UK	United Kingdom
UKB	UK Biobank
WM	White Matter
WMH	White Matter Hyperintensities

PREFACE

The main remit of this work is to make the brain imaging data from UK Biobank a powerful resource for researchers everywhere. With that in mind, I developed and applied an image processing pipeline and a QC framework that automatically generated quality controlled imaging-derived phenotypes (IDPs) and processed images. I also developed a way to unconfound the IDPs, reducing the chances of finding spurious associations between brain imaging variables and other variables in UK Biobank.

To accomplish these goals, I wrote most of the code ($\sim 80\%$) of the UK Biobank processing pipeline, 100% of the code for the QC framework, and $\sim 90\%$ of the code for the confound modelling. Of course, these tools also make use of many other tools developed by other people/teams (e.g. FUNPACK was developed by Paul McCarthy, eddy by Jesper Anderson, AMICO by Alessandro Daducci, etc.).

The writing of this code, and the decision making, was a collaborative process in which the co-authors of each of the main papers (e.g., [Alfaro-Almagro et al. \(2018, 2020\)](#)) contributed in an essential way, but ultimately, the hands-on work, the experiments and the final decisions were mine (in conjunction with my primary supervisor).

I personally checked more than 25,000 T1 images in detail by eye. The first 10,000 were checked as part of the autoQC development and testing, as described in Section 3.3.1. The other 15,000 were checked as part of the resulting QC framework.

The list of papers I have authored or co-authored during this DPhil is listed below; material in this thesis linked to these papers is only taken from papers where I was first author (the first two listed).

- **Alfaro-Almagro, F.**, Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., Miller, K. L., and Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166:400–424.
- **Alfaro-Almagro, F.**, McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M.,

- Miller, K. L., Nichols, T. E., and Smith, S. M. (2020). Confound modelling in uk biobank brain imaging. *NeuroImage*, page 117002.
- Armstrong, N.J., Mather, K.A., Sargurupremraj, M., Knol, M.J., Malik, R., Satizabal, C.L., Yanek, L.R., Wen, W., Gudnason, V.G., Dueker, N.D., **Alfaro-Almagro, F.**, and Elliott, L.T., 2020. Common Genetic Variation Indicates Separate Causes for Periventricular and Deep White Matter Hyperintensities. *Stroke*, pp.119.
 - Bastiani, M., Cottaar, M., Fitzgibbon, S. P., Suri, S., **Alfaro-Almagro, F.**, Sotiropoulos, S. N., Jbabdi, S., and Andersson, J. L. R. (2019). Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction. *NeuroImage*, 184:801–812.
 - Bijsterbosch, J., Harrison, S., Duff, E., **Alfaro-Almagro, F.**, Woolrich, M., and Smith, S. (2017). Investigations into within- and between-subject resting-state amplitude variations. *NeuroImage*, 159:57–69.
 - Elliott, L.T., Sharp, K., **Alfaro-Almagro, F.**, Shi, S., Miller, K.L., Douaud, G., Marchini, J. and Smith, S.M., 2018. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726), pp.210-216.
 - Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., **Alfaro-Almagro, F.**, Glasser, M. F., Duff, E. P., Fitzgibbon, S., Westphal, R., Carone, D., Beckmann, C. F., and Smith, S. M. (2017). Hand classification of fMRI ICA noise components. *NeuroImage*, 154:188–205.
 - Gorgolewski, K. J., **Alfaro-Almagro, F.**, Auer, T., Bellec, P., Capota, M., Chakravarty, M. M., Churchill, N. W., Cohen, A. L., Craddock, R. C., Devenyi, G. A., et al. (2017). Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLOS computational biology*, 13(3):e1005209.
 - Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., **Alfaro-Almagro, F.**, Bell, J. D., Boulton, C., Collins, R., Conroy, M. C., et al. (2020). The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11(1):1–12.
 - 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., Griffanti,

- L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536.
- Mollink, J., Smith, S. M., Elliott, L. T., Kleinnijenhuis, M., Hiemstra, M., **Alfaro-Almagro, F.**, Marchini, J., van Cappellen van Walsum, A.-M., Jbabdi, S., and Miller, K. L. (2019). The spatial correspondence and genetic influence of inter-hemispheric connectivity with white matter microstructure. *Nature Neuroscience*, 22(5):809–819.
 - Nobis, L., Manohar, S. G., Smith, S. M., **Alfaro-Almagro, F.**, Jenkinson, M., Mackay, C. E., and Husain, M. (2019). Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical*, 23:101904.
 - Parker Jones, O., **Alfaro-Almagro, F.**, and Jbabdi, S. (2018). An empirical, 21st century evaluation of phrenology. *Cortex*, 106:26–35.
 - Raman, B., Cassar, M.P., Tunnicliffe, E.M., Filippini, N., Griffanti, L., **Alfaro-Almagro, F.**, Okell, T., Sheerin, F., Xie, C., Mahmood, M. and Mózes, F.E., 2020. Medium-term effects of SARS-CoV-2 infection on multiple vital organs, exercise capacity, cognition, quality of life and mental health, post-hospital discharge. *medRxiv*.
 - Smith, S. M., Vidaurre, D., **Alfaro-Almagro, F.**, Nichols, T. E., and Miller, K. L. (2019). Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539.
 - Smith, S. M., Elliott, L. T., **Alfaro-Almagro, F.**, McCarthy, P., Nichols, T. E., Douaud, G., and Miller, K. L. (2020). Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *eLife*, 9:e52677.
 - Smith, S.M., Douaud, G., Chen, W., Hanayik, T., **Alfaro-Almagro, F.**, Sharp, K. and Elliott, L.T., 2020. Enhanced Brain Imaging Genetics in UK Biobank. *bioRxiv*.

- Sundaresan, V., Griffanti, L., Kindalova, P., **Alfaro-Almagro, F.**, Zamboni, G., Rothwell, P. M., Nichols, T. E., and Jenkinson, M. (2019a). Modelling the distribution of white matter hyperintensities due to ageing on mri images using bayesian inference. *NeuroImage*, 185:434–445.
- Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A. J., **Alfaro-Almagro, F.**, Smith, S. M., and Woolrich, M. W. (2018). Discovering dynamic brain networks from big data in rest and task. *NeuroImage*, 180:646–656.
- Wiberg, A., Ng, M., Al Omran, Y., **Alfaro-Almagro, F.**, McCarthy, P., Marchini, J., Bennett, D. L., Smith, S., Douaud, G., and Furniss, D. (2019). Handedness, language areas and neuropsychiatric diseases: insights from brain imaging and genetics. *Brain*, 142(10):2938–2947.
- Zarnani, K., Nichols, T. E., **Alfaro-Almagro, F.**, Fagerlund, B., Lauritzen, M., Rostrup, E., and Smith, S. M. (2019). Discovering markers of healthy aging: a prospective study in a danish male birth cohort. *Aging (Albany NY)*, 11(16):5943.

Chapter 1

Introduction

1.1 General description of UK Biobank

UK Biobank is a prospective cohort epidemiological study of over 500,000 individuals from across the United Kingdom ([Sudlow et al. \(2015\)](#)). Voluntary participants, aged between 40 and 69 when recruited in 2006–2010, were invited to one of 22 centres across the UK, where rich amounts of data were gathered from the participants. This valuable data resource will provide insight into how the health of the UK population develops over many years and will enable researchers to improve the diagnosis and treatment of common diseases which will inevitably occur in subgroups of the population.

According to Table 2 in [Sudlow et al. \(2015\)](#), the data collected at the baseline assessment through questionnaires and interviews included socio-demographic data (e.g. social class, ethnicity, employment status, marital status, education, income, car ownership), family history and early family exposures (family history of major diseases, birth weight, breast-feeding, maternal smoking, childhood body size, residence at birth), psycho-social factors (neurosis, depression (including bipolar spectrum disorder), social support), environment factors (current address, current (or last) occupation, domestic heating and cooking fuel, housing, means of travel, shift work, mobile phone use, sun exposure), lifestyle (smoking, alcohol consumption, physical activity, diet, sleep), health status (medical history, medications, disability hearing, sight, sexual and reproductive history), hearing threshold (speech reception threshold), and cognitive function (e.g. pairs matching, reaction time, prospective memory, fluid intelligence, numeric memory).

Physical measures taken at the same baseline assessment included blood pressure and heart rate (two automated measures, one minute apart), grip strength (left and right-hand grip strength), anthropometrics (standing and sitting height, weight and bio-impedance, hip and waist circumference), spirometry (up to three measures), bone density (calcaneal ultrasound), arterial stiffness (pulse wave velocity) eye examination (refractive index,

intraocular pressure, acuity, retinal photograph, optical coherence tomography), fitness test (cycle ergometry with ECG heart rate monitoring) and biological samples (blood, urine, and saliva).

Most of the data from the baseline assessments are acquired again at each additional appointment, such as the imaging visits. Additionally, some participants have worn accelerometers for a week to provide physical activity data and many of them also completed online questionnaires regarding diet, cognitive function, occupational history, mental well-being, gastrointestinal health and pain.

The full potential of UK Biobank stems from its access to the NHS health records; participants have agreed for their future health outcome information to be linked into the UK Biobank database. Thanks to this, researchers can aim to find novel biomarkers for early diagnosis of different diseases. If we focus on the brain-related diseases, as described in [Sudlow et al. \(2015\)](#), 1,800 of the imaged participants are expected to develop Alzheimer’s disease by 2022, rising to 6,000 by 2027 (diabetes: 8,000 rising to 14,000; stroke: 1,800 to 4,000; Parkinson’s: 1,200 to 2,800).

Despite its great potential, UK Biobank does not fully represent the general population, as has been suggested in ([Swanson, 2012](#)), and shown in ([Fry et al., 2017](#)) and ([Haworth et al., 2019](#)). This has also been acknowledged by UK Biobank organisation¹. This lack of representation means that the UK Biobank tends to be older, and wealthier than the general population with females being more prevalent too. There is also a “healthy volunteer bias”, which means that the rates of obesity, smoking, alcohol consumption, and health conditions are lower than in the general population of the same age. These considerations, while important in other studies, should not affect our work, as this thesis is focused on method development and not on exposure-disease relationships.

In the rest of this introduction (**chapter 1**), we will give an overview of UK Biobank and the different imaging studies it includes, with particular detail on the brain imaging study. We have included sections of literature review throughout the thesis (primarily towards the start of each core chapter), as we felt this was more useful than collating all

¹Publications using UK Biobank data are advised to include this statement: “While UKB participants are not representative of the general population (and hence cannot be used to provide representative disease prevalence and incidence rates), valid assessment of exposure-disease relationships are nonetheless widely generalisable and do not require participants to be representative of the population at large.” <http://www.ukbiobank.ac.uk/wp-content/uploads/2017/03/access-matters-representativeness-1.pdf>

of the review material within an introductory chapter.

In **chapter 2**, we describe version 1.0 of “FBP” (FMRIB’s Biobank Pipeline). We also include several investigations relating to acquisition and analysis pipeline decisions (such as B0 shimming approaches and registration method for dMRI data), with analyses designed to evaluate relative merits of different approaches.

Chapter 3 describes the different QC (Quality Control) methods developed to make sure that the publicly distributed data can be used by researchers. We specifically focus on methods to assess the quality of T1w images and their non-linear registration to the MNI atlas, as this step is critical for all the rest of the processing. We also describe how we are performing QC on FreeSurfer outputs.

Dealing with confounds is an essential step in large cohort studies to address problems such as unexplained variance and spurious correlations. In **Chapter 4**, we show how we have approached this problem, and we propose a methodology to deal with confounds, not just in UK Biobank, but in other brain imaging studies.

Chapter 5 describes the impact of our work on the neuroimaging research and also future directions of our work in each of the main areas we have investigated: Processing pipeline, QC, and confound modelling.

This work used imaging and non-imaging data from UKB, accessed under data access application 8107. Chapters 2 and 3 mostly correspond to our paper ([Alfaro-Almagro et al., 2018](#)) and were written using the January 2017 data release ($\sim 10,000$ subjects). Chapter 4 corresponds to our paper ([Alfaro-Almagro et al., 2020](#)) and was written using the January 2020 data release ($\sim 40,000$ subjects).

1.2 Imaging studies

In 2014, UK Biobank began the process of inviting back 100,000 of the original volunteers for brain, heart and body imaging. Imaging data for 48,765 volunteers has already been acquired and made available for further research ([Biobank, 2016, 2017](#)). Due to the broad scope (in population size and number of imaging modalities) of brain imaging acquisitions, UK Biobank will become an important resource for the research community.

Four dedicated imaging centres are equipped with identical MRI scanners (3T Siemens

Skyra with software VD13 and 32-channel receive head coil for brain MRI, and Siemens 1.5 Tesla MAGNETOM Aera scanner with VD13A software and a spine and body flex matrix coil for cardiac and abdominal MRI). As shown in multiple studies ([Focke et al., 2011](#); [Chen et al., 2014](#)), having the same scanner and protocol is advantageous for a multi-site study. The main imaging studies in UK biobank are:

- **Brain MRI:** Full details of the brain MRI study (~35 minutes of acquisition) are given in the next section.
- **Cardiac MRI:** This study (~20 minutes of acquisition) aims to gather structural and functional information from the heart. Several automated processing tools have been developed to produce measures related to cardiovascular diseases, such as ventricular mass, ventricular ejection fraction, atrial volume and aortic stiffness.
- **Abdominal MRI:** This whole-body acquisition (~10 minutes) aims to get body fat composition and distribution information, like visceral obesity. So far, no automated tools have been produced to obtain numeric measures, but some semi-automated tools have been developed to get measures of fat, muscle and organs, like visceral and ectopic fat content, or liver measures.
- **Dual-energy X-ray absorptiometry:** A low-dose X-ray study (~20 minutes of acquisition) aims to capture the bone mineral density and body composition of the participants. These measures are automatically generated by the scanner and do not need deep further processing. These X-rays are acquired from the whole body, femur, lumbar spine, and knees. This allows UKB to measure the bone strength and texture, hip structure, and hip and knee osteoarthritis status of the participants.
- **Carotid ultrasound:** This study (~10 minutes of acquisition) gathers measures of vessel thickness (carotid intima-media thickness (CIMT)), and vessel wall and plaque volume. CIMT measures are automatically generated by the acquisition tool and do not need deep further processing. Some automated tools are being developed to generate the vessel wall volume and plaque volume.
- **12-lead electrocardiogram:** These acquisitions (taking less than 1 minute) aim to provide information about heart rhythm and electrical cardiac activity, to find cardiac abnormalities. The ECG system automatically generates a set of measures, and it includes an arrhythmia detection and measures of electrical activity (PR interval, QRS time, and QT interval).

As can be seen, due to the UKB size, many new methods need to be developed to automatically or semi-automatically extract measures from the different modalities that can be used by researchers. For a thorough overview of the imaging enhancement project, see (Littlejohns et al., 2020).

1.3 Brain imaging study

The background to the UK Biobank brain imaging study has been described in detail in Miller et al. (2016), a resource paper that also provides an introduction to the types of biological information that can be extracted from the different brain imaging modalities, as well as example results from early analyses from the first 5,000 subjects' data, and detailed discussions of the challenges of analysing and interpreting such datasets. The data is available online².

In contrast, in this thesis, we focus more specifically on the methodological advances and implementation details in the image processing, QC pipeline, and unconfounding methods that we have developed for UK Biobank brain imaging.

The included modalities are T1 (Processing pipeline in Section 2.2.2), T2 FLAIR (Processing pipeline in Section 2.2.3), swMRI (susceptibility-weighted MRI, Processing pipeline in Section 2.2.4), dMRI (diffusion MRI, Processing pipeline in Section 2.2.7), rfMRI (resting-state functional MRI, Processing pipeline in Section 2.2.8) and tfMRI (task fMRI, Processing pipeline in Section 2.2.9). Section 1.3.1 briefly describes the acquisition protocol and the acquired modalities.

Other large-scale imaging projects include the Lothian birth cohort with 650 participants (Wardlaw et al., 2011), the HCP (Human Connectome Project) with 1,200 participants (Van Essen et al., 2013), the ADNI (Alzheimer's Disease Neuroimaging Initiative) Study with 1,800 participants (Mueller et al., 2005), the ABCD (Adolescent Brain Cognitive Development) Study with 10,000 participants (Casey et al., 2018), the Maastricht Study with 10,000 subjects (Schram et al. (2014)), the German National Cohort with 30,000 subjects (GNC (2014)), and the Rhineland Study with 30,000 subjects (Breteler et al., 2014). More comprehensive lists of population-level brain MRI studies can be found elsewhere³

²<https://amsportal.ukbiobank.ac.uk/sitepages/sign%20in.aspx>

³<http://populationimaging.eu/europeancohorts/>

A different approach to using large amounts of brain MRI data can be found in the ENIGMA consortium, which is gathering data from tens of thousands of participants by combining different studies (including UK Biobank) using advanced meta-analysis techniques ([Bearden and Thompson, 2017](#)).

1.3.1 Acquisition overview

The scanners operate seven days per week, each scanning 18 subjects per day. These acquisition rates demand an extremely time-efficient protocol - each minute of scanning (added into the protocol) in effect costs $\sim$ £1 million, so scanning time has been a key criterion in the protocol development. The total scanning time has been established as 31 minutes (plus 5 minutes for subject adjustment, shimming, etc.). The optimised order of acquisition is (1) T1, (2) resting fMRI, (3) T2 FLAIR, (4) dMRI, (5) swMRI (6) task fMRI, ([Biobank, 2014](#)).

During the acquisition process, if a significant artefact is observed while scanning, and there is enough time to restart the sequence, then the acquisition may be repeated. The most common reason for these artefacts is excessive head movement.

Likewise, as mentioned in [Miller et al. \(2016\)](#), in case of a possible health-related incidental finding noted by the radiographer during the imaging process, a formal radiologist review is undertaken and, if it is confirmed as potentially serious, feedback is given to the participant and their doctor.

To date, our processing pipeline has been applied to 45,483 UK Biobank volunteers scanned from 2014 to 2020 at the four imaging centres. 3,159 of those have gone through a second imaging visit, raising the amount of processed datasets to 48,645 (5,070 of which have not been publicly released yet). The main acquisition parameters for each modality are listed in Table [1.1](#).

A usable T1-weighted image was reported to be obtained for 98% of the first $\sim$ 6,000 subjects [Miller et al. \(2016\)](#). The same figure was reported for the second release (total of $\sim$ 10,000 subjects) in ([Alfaro-Almagro et al., 2018](#)). This number was slightly reduced in later releases, being 95% (41,186 out of 43,575 datasets). These high numbers are an essential point, as a T1 image is a crucial part of the processing pipeline, being essential for making full use of data from the other modalities. At the initial stage, an image was

Table 1.1: UK Biobank brain MRI protocol (31 minutes total scan time).

Modality	Time	Voxel, Matrix	Key Parameters	Volumes
T1	4:54	$1 \times 1 \times 1\text{mm}$ $208 \times 256 \times 256$	3D MPRAGE, sagittal, R=2, TI/TR=880/2000 ms	1
T2 FLAIR	5:52	$1.05 \times 1.0 \times 1.0\text{mm}$ $192 \times 256 \times 256$	FLAIR, 3D SPACE, sagittal, R=2, PF 7/8, fat sat, TI/TR=1800/5000 ms, elliptical	1
swMRI	2:34	$0.8 \times 0.8 \times 3.0\text{mm}$ $256 \times 288 \times 48$	3D GRE, axial, R=2, PF 7/8 TE1/TE2/TR=9.4/20/27ms	2
dMRI	7:08	$2.0 \times 2.0 \times 2.0\text{mm}$ $104 \times 104 \times 72$	MB=3, R=1, TE/TR=92/3600 ms, PF 6/8, fat sat, $b=0 \text{ s/mm}^2$ (5x + 3x phase-encoding reversed), $b=1000 \text{ s/mm}^2$ (50x), $b=2000 \text{ s/mm}^2$ (50x)	105 + 6 (AP + PA)
rfMRI	6:10	$2.4 \times 2.4 \times 2.4\text{mm}$ $88 \times 88 \times 64$	TE/TR=39/735 ms, MB=8, R=1, flip angle 52° , fat sat	490
tfMRI	4:13	$2.4 \times 2.4 \times 2.4\text{mm}$ $88 \times 88 \times 64$	Acquisition same as rfMRI. Task is faces/shapes “emotion” task.	332

R = in-plane acceleration factor, MB = multiband factor, PF=partial Fourier. All non-EPI scans are pre-scan normalised (on-scanner bias-field corrected). Gradient distortion correction is deselected on the scanner and applied in post-processing. SNR, tSNR and motion plots are shown in figures A.1 and A.2 in the Appendix.

deemed unusable only after careful manual quality checks across all modalities⁴. A set of manual classification categories, specific to each modality, were assigned to every image; these categories are described in Section 3.3.1.1. Data for all modalities were acquired and usable in 33,155 datasets (76.09%). These low numbers are due to a lower acquisition rate of task fMRI (the modality that is acquired last and therefore, skipped sometimes to avoid delays). If we do not consider that modality, we have a “usability” rate of 83.57% (36,414 datasets).

In the next two sections, we describe the different modalities that are acquired as part of the brain imaging study and the evolution of the brain imaging protocol.

1.3.2 Acquired modalities overview

T1-weighted structural imaging (“T1”) provides information relating to volumes and morphology of brain tissues and structures. It is also critical for calculations of cross-subject and cross-modality alignments, needed in order to process all other brain modalities. Acquisition is 1mm isotropic resolution using a 3D MPRAGE acquisition. The superior-inferior field-of-view is large (256mm), at little cost, in order to include reasonable amounts of neck/mouth, as those areas hold interest for some researchers.

T2-weighted FLAIR imaging (“T2 FLAIR”) is a structural technique with contrast dominated by signal decay from interactions between water molecules (T2 relaxation times). T2 images depict alterations to tissue properties typically associated with specific pathologies (e.g., white matter lesions). In this modality, the signal from fluid (CSF) is suppressed, and hence it appears dark (unlike either PD or T2w images).

After early piloting, a standard **T2/PD-weighted acquisition** was dropped due to a combination of factors such as overall value and timing practicalities. However, the T2-weighted FLAIR image is still acquired, which is generally of good quality and which shows a strong contrast for white matter hyperintensities.

Susceptibility-weighted imaging (“swMRI”) is a structural technique that is sensitive to “magnetic” tissue constituents (magnetic susceptibility). Data from one scan (including phase and magnitude images from two echo times) can be processed in multiple ways to reflect venous vasculature, microbleeds, quantitative susceptibility mapping

⁴Manual QC was possible because, at that stage, we were dealing with the initial phase (10% of final dataset size)

or aspects of microstructure (e.g., iron, calcium and myelin). Voxel size for this modality is $0.8 \times 0.8 \times 3.0\text{mm}$. Anisotropic voxels enhance contrast in the signal magnitude from sources of signal dephasing, such as iron in veins or microbleeds, but are less ideal for other susceptibility-based processing. Ultimately, however, the decision to acquire anisotropic voxels was motivated by the desire for whole-brain coverage in the face of very limited scan time (2.5 minutes).

Diffusion-weighted imaging (“dMRI”) measures the ability of water molecules to move within their local tissue environment. Water diffusion is measured along a range of orientations, providing two types of useful information. Local (voxel-wise) estimates of diffusion properties reflect the integrity of microstructural tissue compartments (e.g., diffusion tensor and NODDI measures). Long-range estimates based on tract-tracing (tractography) reflect structural connectivity between pairs of brain regions. For the two diffusion-weighted shells, 50 diffusion-encoding directions were acquired (with all 100 directions being distinct). Acquisition phase-encoding direction for the dMRI data is AP (Anterior-to-Posterior). In order to generate appropriate fieldmaps to carry out geometric distortion correction for EPI images (See Section 2.2.5), a reverse (PA: Posterior-to-Anterior) phase-encoding dMRI is also acquired.

Resting-state functional MRI (“rfMRI”) measures changes in blood oxygenation associated with intrinsic brain activity (i.e., in the absence of an explicit task or sensory stimulus). It can generate valuable estimations of the apparent connectivity between pairs of brain regions, as reflected in the presence of spontaneous co-fluctuations in the signal. As implemented in the CMRR multiband acquisition (Moeller et al., 2010), a separate “single-band reference scan” (SBRef) is also acquired. This has the same geometry (including EPI distortion) as the time-series data, but is not part of a low-TR time-series, and hence, without the corresponding T1 saturation effects, has higher between-tissue contrast; this is used as the reference scan in head motion correction and alignment to other modalities. In cases where the SBRef image is missing, the alignments use an average image generated from the first five volumes from the fMRI time-series and selecting the one that correlates most highly with the others (after co-alignment). This procedure is similar to the selection of the best $b=0$ image described in Section 2.2.5.

Task functional MRI (“tfMRI”) uses the same measurement technique as resting-

state fMRI, while the subject performs a particular task or experiences a sensory stimulus. The task used is the Hariri faces/shapes “emotion” task ([Hariri et al., 2002](#); [Barch et al., 2013](#)), as implemented in the HCP, but with shorter overall duration and hence fewer total stimulus block repeats. The participants are presented with blocks of trials and asked to decide either which of two faces presented on the bottom of the screen match the face at the top of the screen, or which of two shapes presented at the bottom of the screen match the shape at the top of the screen. The faces have either angry or fearful expressions. This task was chosen to engage a range of sensory, motor and cognitive systems.

1.3.3 Protocol evolution

Throughout the development of the Biobank imaging procedures, the imaging acquisition has been divided into six protocol phases, with each phase corresponding to (generally minor, very early) adjustments in the acquisition protocol.

The original aim in the UK Biobank brain imaging component was to maintain a fixed acquisition protocol during the 5-6 years that the scanning will require, or, at least, to have maximum compatibility with eventual choices. However, very early improvements in the dMRI and T2 FLAIR protocols were found to be valuable, resulting in large enough data improvements to outweigh the priority of keeping the protocol fixed (and taking into account the relatively small numbers of datasets affected). This change was made at the beginning of protocol “Phase 3”, hence affecting just ~ 550 datasets. In [Section A.2](#), we describe one protocol change for which we carried out specific custom analyses to help verify the protocol change decision (relating to field shimming). The different phases are described in detail below. The number of processed subjects in each phase can be found in [Table A.5](#) from the Appendix.

There also have been some minor software upgrades (for obvious reasons the imaging study seeks to minimise any significant hardware or software changes). These upgrades and the number of subjects belonging to each of them can be found in [Tables A.3](#) and [A.4](#) from the Appendix. As shown in [Chapter 4.3.3](#), these software upgrades are not significant (they explain a negligible amount of variance of the data, and they do not affect the correlations between IDPs and non-IDPs in a meaningful way).

Protocol Phase 1 - Up to 08/05/2014

Only ten datasets were acquired within this initial phase, after which several major improvements were made in the protocol. These data sets have not been included in the imaging and IDP data sets that have been released.

Protocol Phase 2 - 09/05/2014 to 11/08/2014

These datasets were acquired with a different protocol from Phase 1. One of the main features of this phase is the fact that all on-scanner gradient distortion correction was turned off, as this does not work well in 3D for EPI acquisitions (thus, this step is done inside the processing pipeline, as described in Section 2.2.1).

A further difference in Phase 2 scans which does not present incompatibility problems is that resting fMRI and task fMRI protocols had additional time points (approximately 30s) compared with later scans. Our processing discards these, for compatibility.

A number of improvements in Phase 3 for T2 FLAIR and dMRI (below) make these acquisitions in Phase 2 incompatible with later phases. The raw T2 FLAIR and dMRI NIFTI images from Phase 2 have been made available via the UK Biobank database, but due to these incompatibilities, they have not been used in the full image processing pipeline and were not used in the IDP (Image-Derived Phenotype) generation.

Protocol Phase 3 - 12/08/2014 to 22/09/2014

This phase implemented some important changes. The range of intensity values in the T1 images was increased in the analog-to-digital conversion to make better use of the dynamic range of the 12-bit integer DICOM outputs. Therefore, intensities for this modality should be normalised when combining or comparing subjects from different phases (indeed in general the overall intensity scaling of images like T1 cannot be assumed to be of any given value).

T2 FLAIR switched to using elliptical k-space coverage to reduce acquisition time with no significant loss in image quality and implemented 7/8 partial Fourier, which reduced image blurring with a small time penalty. This was an important change with respect to the previous phase, where Partial Fourier was set to 6/8.

dMRI acquisitions started using “monopolar” diffusion encoding from this phase. The reason for this decision is that this improves data SNR, as explained in (Alfaro-Almagro et al., 2018). Other minor differences with Phase 2 dMRI scans are: one more diffusion

encoding direction per shell (going from 7/49/49 to 8/50/50 for shells $b = 0/1000/2000$), a reduced flip angle (78/160 instead of the former 93/180), to avoid overflipping and improve B1 homogeneity, and a reduced number of reversed encoding direction $b=0$ images for use in EPI distortion correction (3 instead of 5).

Protocol Phase 4 - 23/09/2014 to 26/11/2014

A number of minor changes were made between Phase 3 and 4:

- Three $b=0$ scans were removed in dMRI (Going from 6 to 3).
- The T2 FLAIR was moved, within the protocol, to run after the fMRI scans.
- A new auto-shimming approach was put in place, with a reduced shimming field-of-view and fewer shim iterations. The evaluation of this change can be seen in Section [A.2](#) of the Appendix.

Protocol Phase 5 - 27/11/2014 to 28/07/2016

The only change from Phase 4 to Phase 5 was to upgrade the multiband software from CMRR to version “R012”, a change that has no impact on data acquired with the Biobank protocol (as shown in Section [4.3.3](#)).

Protocol Phase 6 - 29/07/2016 to present

This is the current protocol up to the time of the writing of this thesis. There is only one minor change from Phase 5 to Phase 6: the upgrade on the multiband software from CMRR to version “R014”. As in the previous case, this change did not have any impact on data acquired (again, as shown in Section [4.3.3](#)).

1.3.4 Data release history

There have been five public data release events with the data we produced. These releases happened in this timeline:

- October 2015: First data release with $\sim 5,700$ usable subjects. It corresponds to batch 1 in Table [A.2](#).
- January 2017: Second data release with new $\sim 4,200$ usable subjects, raising the total number to $\sim 10,000$. It corresponds to batch 2 in Table [A.2](#).
- December 2017: Third data release with new $\sim 4,500$ usable subjects, raising the total number to $\sim 14,500$. It corresponds to batch 3 in Table [A.2](#).

- May 2019: Fourth data release with new $\sim 6,900$ usable subjects, raising the total number to $\sim 21,400$. It corresponds to batch 4 in Table [A.2](#).
- January 2020: Fifth data release with new $\sim 16,200$ usable subjects, raising the total number to $\sim 40,000$. It corresponds to batches 4 and 5 in Table [A.2](#). This release includes the second scan visit of $\sim 1,500$ of previously processed subjects.
- Late 2020 (to come shortly): Latest data release, with $\sim 3,500$ new subjects before QC. This includes the second scan visit of $\sim 1,500$ of previously processed subjects. This new release only needs the QC manual inspection described in Section [3.3](#) before being publicly distributed.

Chapter 2

Image Processing and Pipeline for Brain Imaging from UK Biobank

2.1 Introduction: Processing brain imaging data

Brain imaging data is not immediately usable for extracting biologically-meaningful information in its raw form. It needs to be processed using a wide variety of software tools, optimised and combined into a processing pipeline.

There are many excellent examples of processing pipelines in the literature. Tools such as those described in [Wei et al. \(2002\)](#) (segmentation), [Cook et al. \(2006\)](#) (diffusion) or [Song et al. \(2011\)](#) (resting-state), are specific solutions for the processing of specific image modalities. However, for UK Biobank, we need a robust pipeline that can process and integrate across many different functional and structural modalities. The minimal preprocessing pipelines developed by the Human Connectome Project (HCP) ([Glasser et al., 2013](#)) are good examples of pipelines optimised for the specific needs of a project with very special characteristics. Following the same philosophy, we have developed a fully automated processing pipeline primarily based around FSL software ([Jenkinson et al., 2012](#)). In Section 2.2, we describe the pipeline in detail, as well as the choices we have made in its development.

The tests performed in this chapter to evaluate the different steps of the pipeline were performed using randomly selected subjects. We did not attempt to exclude subjects having reported disease, because:

- At the start of the work we did not have access to non-imaging variables (due to the specifics of the access agreement).
- Much of the clinical outcomes data has been rather delayed in getting into the core databases,
- In any case, less than 10% of the participants have mental health or neurological issues diagnosed, as discussed in ([Smith et al., 2019](#)).

- The effects of most neurological and psychiatric diseases on the imaged brain are quite subtle and unlikely to interact with the successful running of our pipelines. The clear exception to this is cases of large structural abnormality (pathology) which we aim to identify through our QC framework.

2.1.1 Imaging-Derived Phenotypes (IDPs)

A goal of UK Biobank is to identify new relationships between different phenotypic features and human diseases, in the hope that they may be used as biomarkers for early diagnosis. Therefore, in addition to preprocessing the brain imaging data (e.g., aligning modalities and removing artefacts), our pipeline also generates more than 5,000 “imaging-derived phenotypes” (IDPs). These IDPs include summary measures such as subcortical structure volumes, microstructural measures in major tracts (DTI and NODDI measures (Zhang et al., 2012)), and structural/functional connectivity metrics. IDPs for the first $\sim 40,000$ subjects have been publicly released¹ and can be used in combination with other non-imaging data from Biobank to identify disease risk factors and biomarkers. Each IDP is presented as a separate data field within the UK Biobank showcase (except for the rfMRI netmats and node amplitudes, which are saved en masse within “bulk” variables in the database). There are 5632 available IDPs in the showcase². This is a detailed description of those IDPs:

- **892 Basic IDPs³**
 - 3 IDPs: Discrepancy between the T1 structural image and the standard (population average) template image (See Section 3.3.1.1).
 - 2 IDPs: Reciprocal of the signal to noise ratio (SNR) and contrast to noise ratio (CNR) of the T1 (See Section 3.3.1.1).
 - 5 IDPs: Cost function for the linear alignment of the 5 modalities to the T1.
 - 2 IDPs: Head motion for resting-state fMRI and task fMRI.
 - 3 IDPs: Reciprocal of the temporal signal to noise ratio (TSNR) for resting-state fMRI, cleaned resting-state fMRI (After running FIX) and task fMRI.
 - 1 IDP: Number of dMRI outlier slices (across all slices in all volumes) detected

¹The raw and processed imaging data, IDPs and non-imaging measures in UK Biobank are available to researchers worldwide following a data access application procedure

²<http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=9028>

³The list and description of these basic IDPs can be found in:
https://git.fmrib.ox.ac.uk/falmagro/UK_biobank_pipeline_v.1/-/blob/master/bb_IDP/list.txt

and corrected with eddy.

- 11 IDPs: SIENAX outputs for normalised and non-normalised volumes (See Section 3.3.1.1).
- 15 IDPs: Volumes of the 15 subcortical structures identified with FIRST (See Section 3.3.1.1).
- 139 IDPs: Volumes in grey matter in 139 cortical structures.
- 1 IDP: Total volume of white matter hyperintensities using BIANCA.
- 14 IDPs: Median T2* intensity (from SWI data) in 7 paired subcortical structures from FIRST.
- 12 IDPs: Median and 90th percentile of the 3 main contrasts in the task fMRI analysis (shapes, faces and faces-shapes) and their corresponding zstats.
- 4 IDPs: Median and 90th percentile of the faces-shapes contrast in the task fMRI analysis in the amygdala and their corresponding zstats.
- 432 IDPs: Mean intensity for dtifit outputs (FA, MD, MO, L1, L2, L3) and NODDI outputs (ICVF, OD, ISOVF) for the 48 tracts segmented with the TBSS-like analysis.
- 243 IDPs: Mean intensity for dtifit outputs (FA, MD, MO, L1, L2, L3) and NODDI outputs (ICVF, OD, ISOVF) for the 27 tracts segmented with the probabilistic tractography analysis.
- 4 IDPs: Center of Gravity of the head inside the scanner and position of the table.
- 1 IDP: Imaging center

- **3,466 Functional Connectivity IDPs**

- 21 IDPs: Amplitudes for the 21 nodes analysed in the functional network connectivity analysis with low dimensionality (25 components before clean-up).
- 210 IDPs: Full normalised temporal correlation between every node for the 21 nodes analysed in the functional network connectivity analysis with low dimensionality (25 components before clean-up).
- 210 IDPs: Partial temporal correlation between every node for the 21 nodes analysed in the functional network connectivity analysis with low dimensionality (25 components before clean-up).
- 55 IDPs: Amplitudes for the 55 nodes analysed in the functional network

connectivity analysis with high dimensionality (100 components before clean-up).

- 1485 IDPs: Full normalised temporal correlation between every node for the 55 nodes analysed in the functional network connectivity analysis with high dimensionality (100 components before clean-up).
- 1485 IDPs: Partial temporal correlation between every node for the 55 nodes analysed in the functional network connectivity analysis with high dimensionality (100 components before clean-up).
- **1,273 FreeSurfer IDPs⁴**
 - 1 IDP: FreeSurfer was run with T1 + T2 Flair or just T1 (True/False)
 - 99 IDPs: Metrics from aseg (volumetric segmentation) (Fischl et al., 2004)
 - * 13 IDPs: Global mean intensity
 - * 22 IDPs: Global volume
 - * 2 IDPs: Global volume ratio
 - * 28 IDPs: Mean intensity (L + R)
 - * 32 IDPs: Volume(L + R)
 - * 2 IDP: Number of holes (L + R)
 - 121 IDPs: Sub-segmentation volumes
 - * 20 IDPs: Volume of Amygdala nuclei (L + R) (Saygin et al., 2017)
 - * 44 IDPs: Volume of Hippocampal subfields (L + R) (Iglesias et al., 2016)
 - * 52 IDPs: Volume of Thalamic nuclei (L + R) (Iglesias et al., 2018)
 - * 5 IDPs: Volume of Brainstem subsegmentations (Iglesias et al., 2015)
 - 268 IDPs: Desikan-Killiany atlas (aparc - pial) (Desikan et al., 2006)
 - * 2 IDPs: Total surface area (L + R)
 - * 2 IDPs: Total thickness (L + R)
 - * 66 IDPs: Surface area of the WM parcellations (L + R)
 - * 66 IDPs: Surface area of the pial parcellations (L + R)
 - * 66 IDPs: Thickness of the parcellations (L + R)
 - * 66 IDPs: Volume of the parcellations (L + R)
 - 70 IDPs: White-gray contrast atlas (w-g.pct) (Salat et al., 2009)
 - * 70 IDPs: Grey-white intensity contrast ratio of the parcellations (L + R)

⁴The list and description of these basic IDPs can be found in:
https://git.fmrib.ox.ac.uk/falmagro/UK_biobank_pipeline_v.1/-/blob/master/bb_data/FS_headers.info.txt

- 84 IDPs: Brodmann Areas atlas (BA_exvivo) (Fischl et al., 2008)
 - * 28 IDPs: Surface area of the parcellations (L + R)
 - * 28 IDPs: Thickness of the parcellations (L + R)
 - * 28 IDPs: Volume of the parcellations (L + R)
- 186 IDPs: Desikan-Killiany-Tourville atlas (DKTatlas) (Klein and Tourville, 2012)
 - * 62 IDPs: Surface area of the parcellations (L + R)
 - * 62 IDPs: Thickness of the parcellations (L + R)
 - * 62 IDPs: Volume of the parcellations (L + R)
- 444 IDPs: Destrieux atlas (a2009s) (Destrieux et al., 2010)
 - * 128 IDPs: Surface area of the parcellations (L + R)
 - * 128 IDPs: Thickness of the parcellations (L + R)
 - * 128 IDPs: Volume of the parcellations (L + R)

2.2 Automated processing pipelines

All processing was performed on a dedicated cluster composed of CPU servers (typical speed 2.66 GHz per core, average RAM 10 GB per core) as well as GPU servers for the GPU-optimised parts of the processing pipeline (Nvidia Tesla K40/K80 GPUs for processing with Cuda 5.5⁵).

Image reconstruction from k-space is performed on the scanner using standard Siemens reconstructions, except for multiband EPI (reconstructed using custom CMRR code). Partial bias field correction in structural data is performed via the on-scanner “pre-scan normalise” option. No on-scanner gradient distortion correction is applied. After reconstruction and on-scanner preprocessing, image data is received (in DICOM format) by the pipeline. An overview of the processing pipeline can be seen in Figure 2.2.

The source code for the pipeline can be found online⁶.

2.2.1 Initial preprocessing steps

Before any modality-specific processing, there are a series of steps that need to be performed for every modality. Those steps include: Conversion from DICOM to NIFTI,

⁵GPU processing is used for the FSL eddy and bedpostx tools, with additional tools being ported to GPU in the future

⁶https://git.fmrib.ox.ac.uk/falmagro/UK_biobank_pipeline_v_1

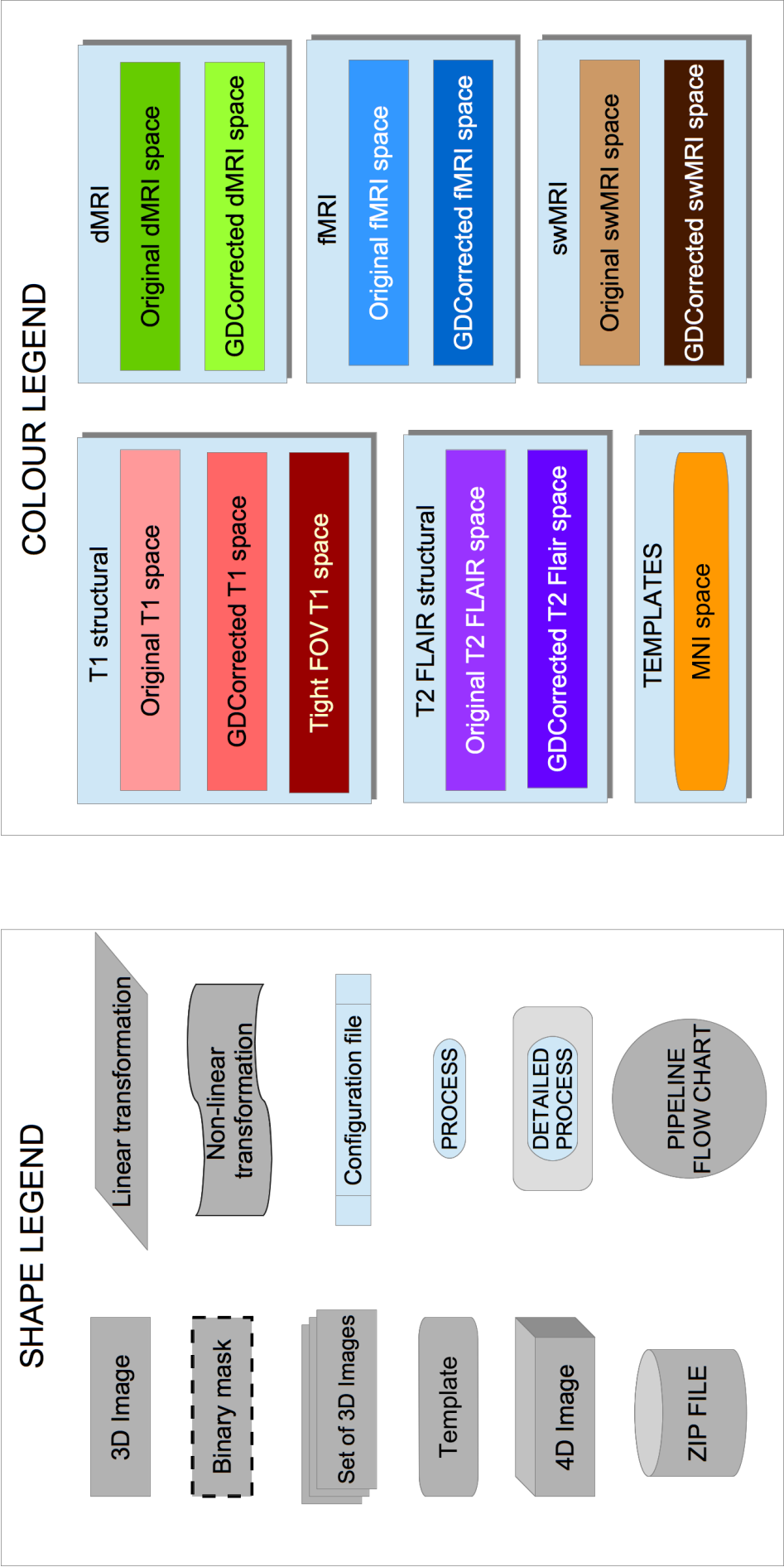


Figure 2.1: Flowcharts legend. All linear registrations shown in the pipeline are rigid body transformations (6 degrees of freedom) in intra-subject operations and affine transformations (12 degrees of freedom) when registering to MNI template.

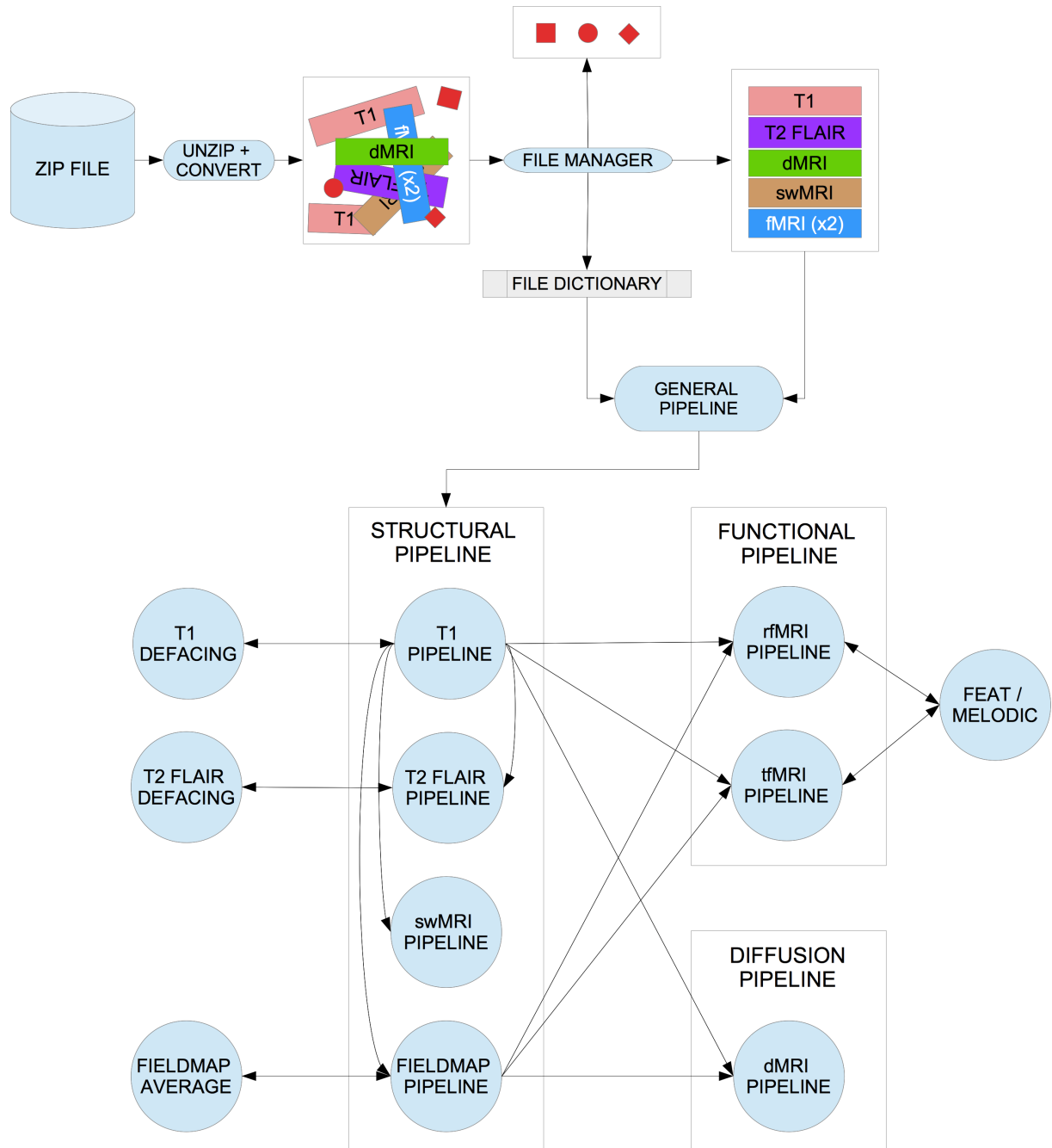


Figure 2.2: General flowchart for the pipeline.

organisation of the data into a standardised file/directory structure, image anonymisation (where necessary), and gradient distortion correction.

2.2.1.1 Data conversion: DICOM to NIFTI

All DICOM image files are converted to NIFTI format using `dcm2niix` (Li et al., 2016). This tool also generates the diffusion-encoding b-value and b-vector files, as well as a JSON file for each NIFTI image with meta-information such as: acquisition date and time, echo time, repetition time, effective echo spacing, encoding direction, magnetic field strength, flip angle and normalisation by the scanner.

Image data is available from UK Biobank in both DICOM and (separately) NIFTI formats. Both forms include the raw (non-processed) image-space data, the only differences being that the NIFTI-version T1/T2 structural images are defaced for subject anonymity (as described below), and the multi-coil (pre-combination) swMRI data is only available in the DICOM downloads. Researchers download individual zip files corresponding to one modality from one subject for one data format (DICOM or NIFTI).

The NIFTI versions are the recommended option, partly because for each modality a small number of simple and consistently named images are provided (e.g., “T1”, “rfMRI”), as opposed to thousands of separate DICOM files (with complex naming and somewhat variable conventions dictated by the scanner). Also, the NIFTI downloads, while overall only being 40% larger than pure-raw DICOM downloads, include not just the raw images, but also images output by the processing pipeline, for example after gradient distortion correction (for all modalities), and correction for eddy currents and head motion (dMRI data), and artefact removal (rfMRI data).

2.2.1.2 Data organization

The directory tree structure of the processed data is available online⁷.

For each subject, the raw and processed imaging data files are automatically organised into subfolders according to the different modalities, with a uniform naming scheme for files and directories. To achieve this, we need to check all the files that have been converted from DICOM to NIFTI, find the ones that correspond to each modality (by matching a set of naming patterns), sort out the possible problems with the number of files found

⁷<http://www.fmrib.ox.ac.uk/ukbiobank>

per modality (missing files or having more files than expected) and finally, renaming the result according to the naming scheme. To resolve problems regarding the number of found files, we developed the next algorithmic rules:

Algorithm 1 Logic to manage swMRI images

Require: swMRI images, PHASE TE1 directory, PHASE TE2 directory

Require: MAGNITUDE TE1 directory, MAGNITUDE TE2 directory

Require: swMRI directory, swMRI unclassified directory

```

if (num(swMRI images) = 134) then
  for all swMRI in swMRIimages do
    if swMRI is coil image then
      swMRI  $\Rightarrow$  coilimages
    else
      swMRI  $\Rightarrow$  mainImages
    end if
  end for
  for all swMRI in coilimages do
    if swMRI is phase image then
      if TE(swMRI) = TE1 then
        swMRI  $\Rightarrow$  PHASE TE1 directory
      else
        swMRI  $\Rightarrow$  PHASE TE2 directory
      end if
    else
      if TE(swMRI) = TE1 then
        swMRI  $\Rightarrow$  MAGNITUDE TE1 directory
      else
        swMRI  $\Rightarrow$  MAGNITUDE TE2 directory
      end if
    end if
  end for{Finish with Coil Images}
  if num(swMRI Phase TE1) in mainImages = 1 then
    swMRI Phase TE1  $\Rightarrow$  swMRI directory
  else if num(swMRI Magnitude TE1) in mainImages = 2 then
    swMRI Magnitude TE1  $\Rightarrow$  swMRI directory
  else if num(swMRI Phase TE2) in mainImages = 1 then
    swMRI Phase TE2  $\Rightarrow$  swMRI directory
  else if num(swMRI Magnitude TE2) in mainImages = 2 then
    swMRI Magnitude TE2  $\Rightarrow$  swMRI directory
  end if{Finish with Main Images}
  for all swMRI not classified do
    swMRI  $\Rightarrow$  swMRI unclassified directory
  end for
end if

```

Algorithm 2 Logic to manage T1w images

Require: T1 images, T1 directory, unclassified directory, raw directory.

{Same rules apply to T2 FLAIR.}

```

1: if (num(T1) >= 2) then
2:   if (num(T1) = 2) then
3:     Normalised T1  $\Rightarrow$  T1 directory {Check JSON file}
4:     Not normalised T1  $\Rightarrow$  raw directory {Check JSON file}
5:   else
6:     Last Normalised T1  $\Rightarrow$  T1 directory {In acquisition order}
7:     Last not Normalised T1  $\Rightarrow$  raw directory {In acquisition order}
8:     Rest of T1s  $\Rightarrow$  unclassified directory {If any}
9:   end if
10: end if

```

Algorithm 3 Logic to manage fMRI images

Require: fMRI images, fMRI directory, unclassified directory,.

{Same rules apply to rfMRI and tfMRI.}

```

1: if (num(fMRI images) = 1) then
2:   if (fMRI image is 4D) then
3:     generate SBRef(fMRI image) {The image did not have SBRef}
4:     fMRI image + SBRef  $\Rightarrow$  fMRI directory
5:   end if
6: else if (num(fMRI images) = 2) then
7:   BiggestImage  $\leftarrow$  max(dim(fMRI images)) {Find the biggest image}
8:   if BiggestImage is 4D then
9:     if SmallestImage is 3D then
10:      SBRef  $\leftarrow$  SmallestImage
11:    else
12:      generate SBRef(BiggestImage) {The image did not have SBRef}
13:    end if
14:    fMRI image + SBRef  $\Rightarrow$  fMRI directory
15:    Rest of fMRIs  $\Rightarrow$  unclassified directory {If any}
16:  end if
17: else
18:   BiggestImage  $\leftarrow$  max(dim(fMRI images)) {Find the biggest image}
19:   if BiggestImage is 4D then
20:     if Previous to BiggestImage is 3D then
21:       SBRef  $\leftarrow$  Previous to BiggestImage
22:     else
23:       generate SBRef(BiggestImage) {The image did not have SBRef}
24:     end if
25:     fMRI image + SBRef  $\Rightarrow$  fMRI directory
26:     Rest of fMRIs  $\Rightarrow$  unclassified directory {If any}
27:   end if
28: end if

```

Algorithm 4 Logic to manage dMRI images

```

Require: dMRI images, dMRI directory, unclassified directory
for all dMRI in dMRI images do
  if EncodingDirection(dMRI) = AP then
    dMRI  $\Rightarrow$  APs
  else
    dMRI  $\Rightarrow$  PAs
  end if
end for
for group in (APs, PAs) do
  if (num(group)  $\geq$  1) then
    BiggestImage  $\leftarrow$  max(dim(group)) {Find the biggest image}
    BiggestImage + bval + bvec  $\Rightarrow$  dMRI directory
    Rest of dMRIs in group  $\Rightarrow$  unclassified directory {If any}
  end if
end for

```

This file organisation can be converted to BIDS file-naming format ([Gorgolewski et al., 2016](#)) with a conversion script included in the pipeline.

2.2.1.3 Image anonymisation

To protect study participant anonymity, the header data does not contain sensitive information such as name and any other information that could be used to identify the participant.

Furthermore, the high-resolution structural images (T1 and T2) are automatically “defaced” by masking out voxels in the face and ear regions. This is accomplished by estimating a linear transformation between the original data and a standard coordinate system (an expanded MNI152 template space)⁸, and back-projecting a standard-space mask of the face and ear regions into the native data. This process is depicted in [Figure 2.3](#). To ensure that this process did not unduly remove within-brain voxels, we calculated the overlap between non-defaced brain masks and the defacing masks. Overlap was present in only 0.5% of the subjects⁹ and, within these subjects, there was nearly no loss of brain voxels (worst case having 68 overlap voxels out of 1725983 brain voxels).

This defacing procedure is similar to common practice such as in the HCP. The raw, non-defaced DICOM T1 and T2 data is classified as “sensitive” by UK Biobank; researchers

⁸Section [2.2.2](#)

⁹33 out of 5822

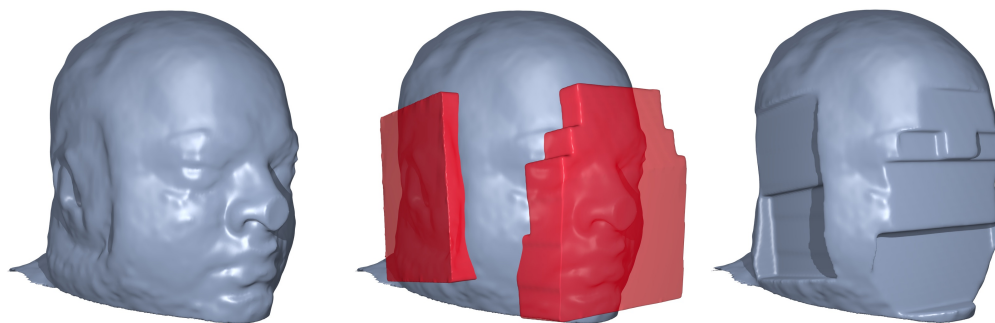


Figure 2.3: Example of defacing. *Left:* Original T1 volume (non-Biobank subject). *Center:* Applying defacing masks. *Right:* Defaced T1 volume.

requiring raw DICOM non-defaced T1/T2 data should contact UK Biobank to discuss special access requirements.

2.2.1.4 Gradient distortion correction

Full 3D gradient distortion correction (GDC) is not available on the scanner for EPI data (gradient distortions are only corrected within-plane for 2D EPI data), so GDC is entirely applied within the processing pipeline.

GDC is a necessary step, as shown in Figure 2.4. Tools developed by the FreeSurfer and HCP teams are used for applying the correction, available through the HCP GitHub account¹⁰. Running these tools also requires a proprietary Siemens data file that describes the gradient nonlinearities¹¹.

2.2.2 T1 pipeline

T1 structural images are used as a reference for all other modalities. The processing performed on these images should avoid any kind of unnecessary smoothing or interpolation. Therefore, all linear and non-linear pipeline transformations are estimated, but not applied until as late as possible in the processing pipeline when all transforms can be combined and applied with a single interpolation.

As can be seen in Figure 2.5, the pipeline runs a GDC, cuts down the field of view (FOV), calculates a registration (linear and then non-linear) to the standard atlas, applies brain extraction, performs defacing and finally segments the brain into different tissues and subcortical structures.

¹⁰<https://github.com/Washington-University/Pipelines>

¹¹More information about this file can be found in:
<https://github.com/Washington-University/Pipelines/wiki/FAQ>

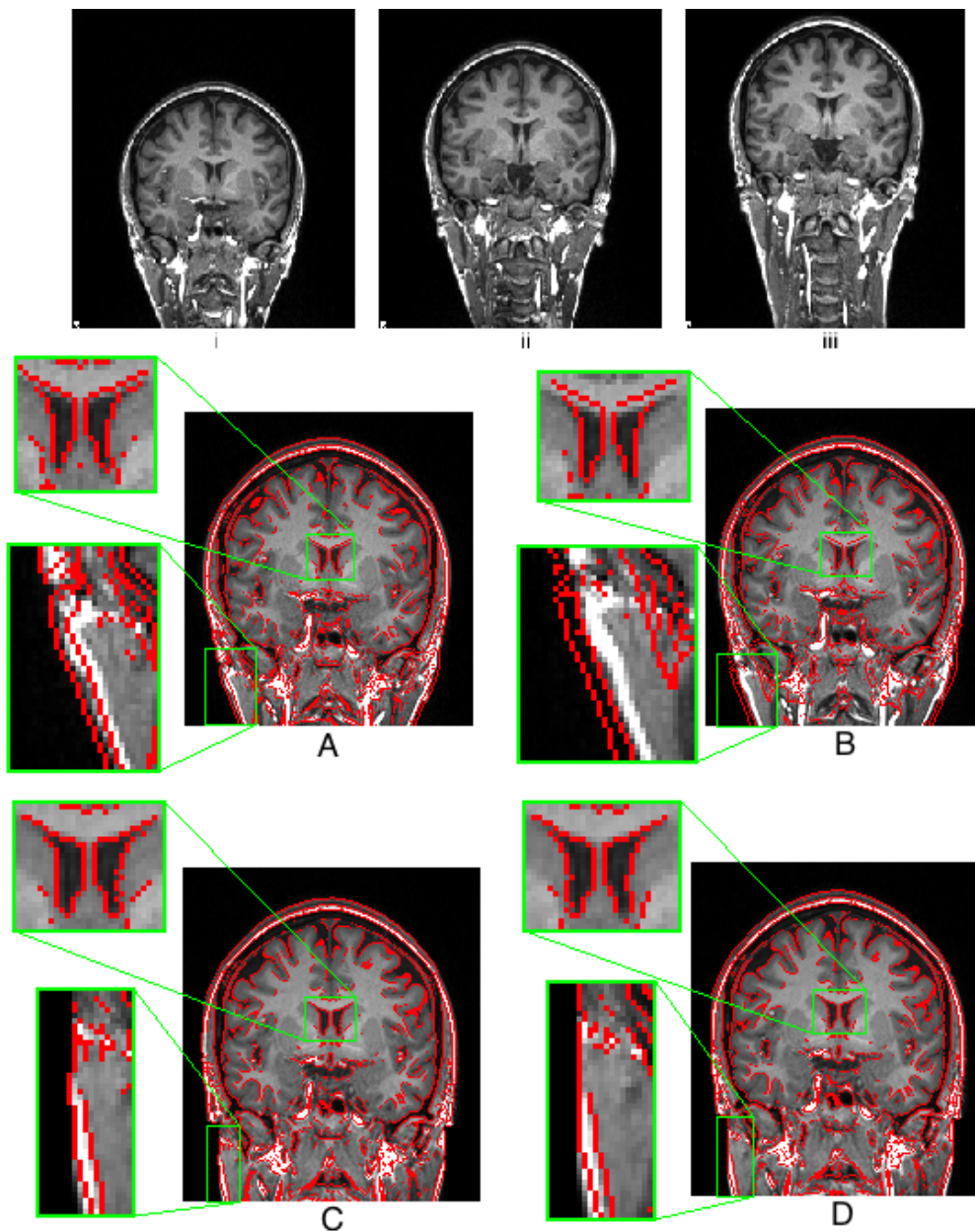


Figure 2.4: One volunteer was scanned for testing purposes at different positions on the scanning table, depicting the effect of gradient distortion (note the warping in the neck). Correction for gradient distortion (*C* and *D*) substantially improves alignment. *i*: Subject placed low in the scanner (Baseline). *ii*: Subject placed in the centre of the scanner. *iii*: Subject placed higher into the scanner. *A*: *ii* (red outline) linearly registered to *i* (background) (Cross correlation: 0.90). *B*: *iii* (red outline) linearly registered to *i* (background) (Cross correlation: 0.83). *C*: *ii* (red outline) linearly registered to *i* (background) after GDC (Cross correlation: 0.98). *D*: *iii* (red outline) linearly registered to *i* (background) after GDC (Cross correlation: 0.96). Regions highlighting the largest effect (i.e. improved alignment of the ventricle) of the GDC have been zoomed to a 3:1 scale.



Figure 2.5: Flowchart for the T1 processing pipeline.

In more detail: the pipeline first applies GDC for the original T1 image as described above. The FOV is then cut down to reduce the amount of non-brain tissue (primarily empty space above the head and tissue below the brain) to improve robustness and accuracy of subsequent registrations. Tools used to achieve this include BET (Brain Extraction Tool, [Smith \(2002\)](#)) and FLIRT (FMRIB’s Linear Image Registration Tool, [Jenkinson and Smith \(2001\)](#); [Jenkinson et al. \(2002\)](#))¹², in conjunction with the MNI152 “non-linear 6th generation” standard-space T1 template¹³. This results in a reduced-FOV T1 head image.

The next step is a non-linear registration to MNI152 space using FNIRT¹⁴ (FMRIB’s Nonlinear Image Registration Tool [Andersson et al. \(2007b,a\)](#)). This is a critical step in the pipeline, as many subsequent processing steps depend on accurate registration to standard space. We used the 1mm resolution version of the MNI152 template as the reference space.

A particular issue for the non-linear registration was the fact that T1 images in UK Biobank have brighter internal carotid arteries than those in the MNI152 template. This affects the non-linear registration procedure, resulting in distortion in the temporal lobes. Hence, we created a custom reference brain mask to exclude this part of the image when estimating the transformation (see Figure 2.6).

¹²Default cost function: Correlation ratio

¹³<http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152Nlin6>

¹⁴Configuration file for optimal registration is distributed with the pipeline code

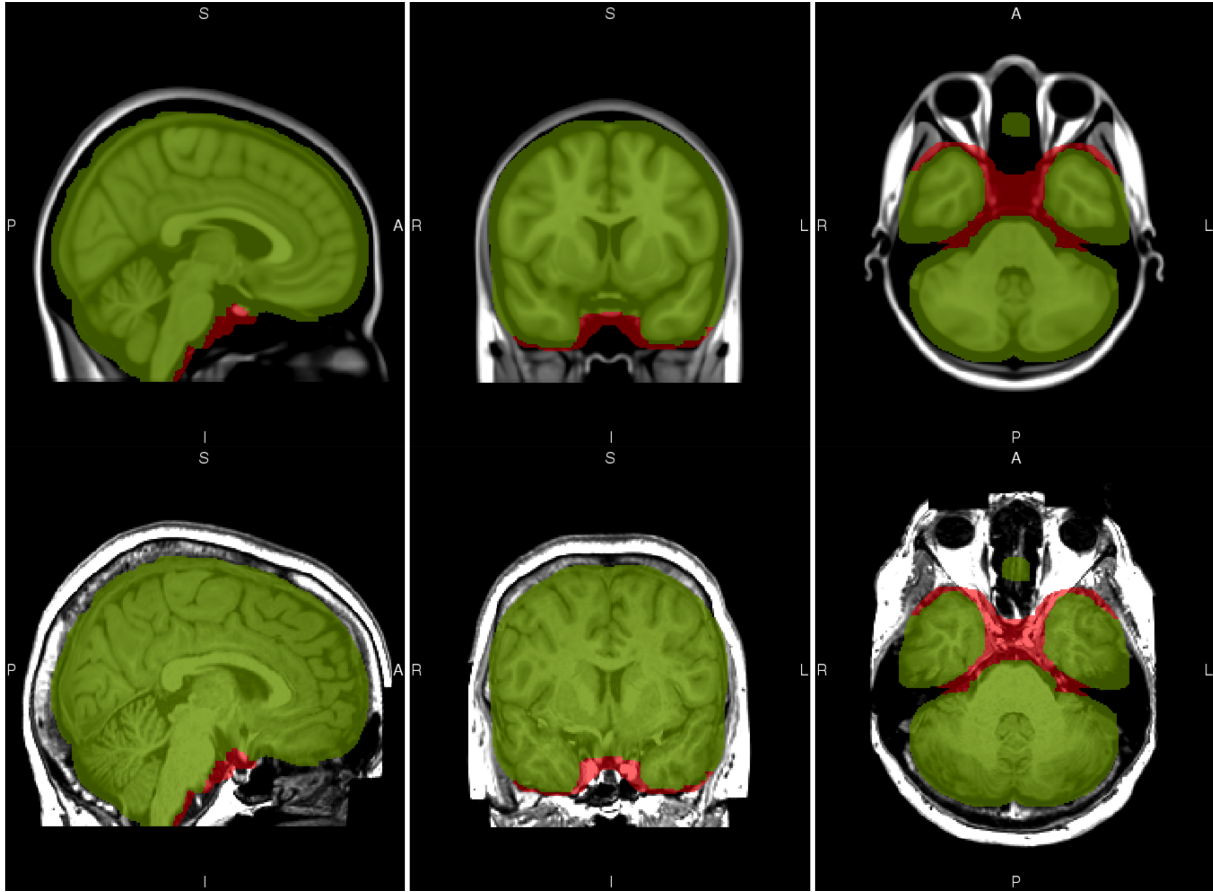


Figure 2.6: The modified dilated brain mask used to improve non-linear registration. *Top*: MNI template with the two brain masks. The red mask is the original brain mask. The green mask is the modified version for UK Biobank. *Bottom*: Subject registered to the MNI template with the two brain masks. The red mask is the original brain mask. The green mask is the modified version for UK Biobank.

All of the transformations estimated above (GDC, linear and non-linear transformations to MNI152) are then combined into one single non-linear transformation, which allows the original T1 to be transformed into MNI152 space (or vice versa) in a single step.

Using the inverse of the MNI152 alignment warp, a standard-space brain mask is transformed into the native T1 space and applied to the T1 image to generate a brain-extracted T1; this brain extraction replaces the earlier brain-extraction output created by BET (which was needed just for the initial registration stages). Similarly, as discussed above, a defacing mask (defined as a set of boxes removing eyes, nose, mouth and ears) is transformed into T1 space using the linear transformation and applied for anonymisation purposes. Only the linear transformation is used for defacing, as non-linear registration is performed on brain-extracted images, meaning that the warp field in out-of-brain regions is poorly conditioned. This process is shown in Figure 2.7.

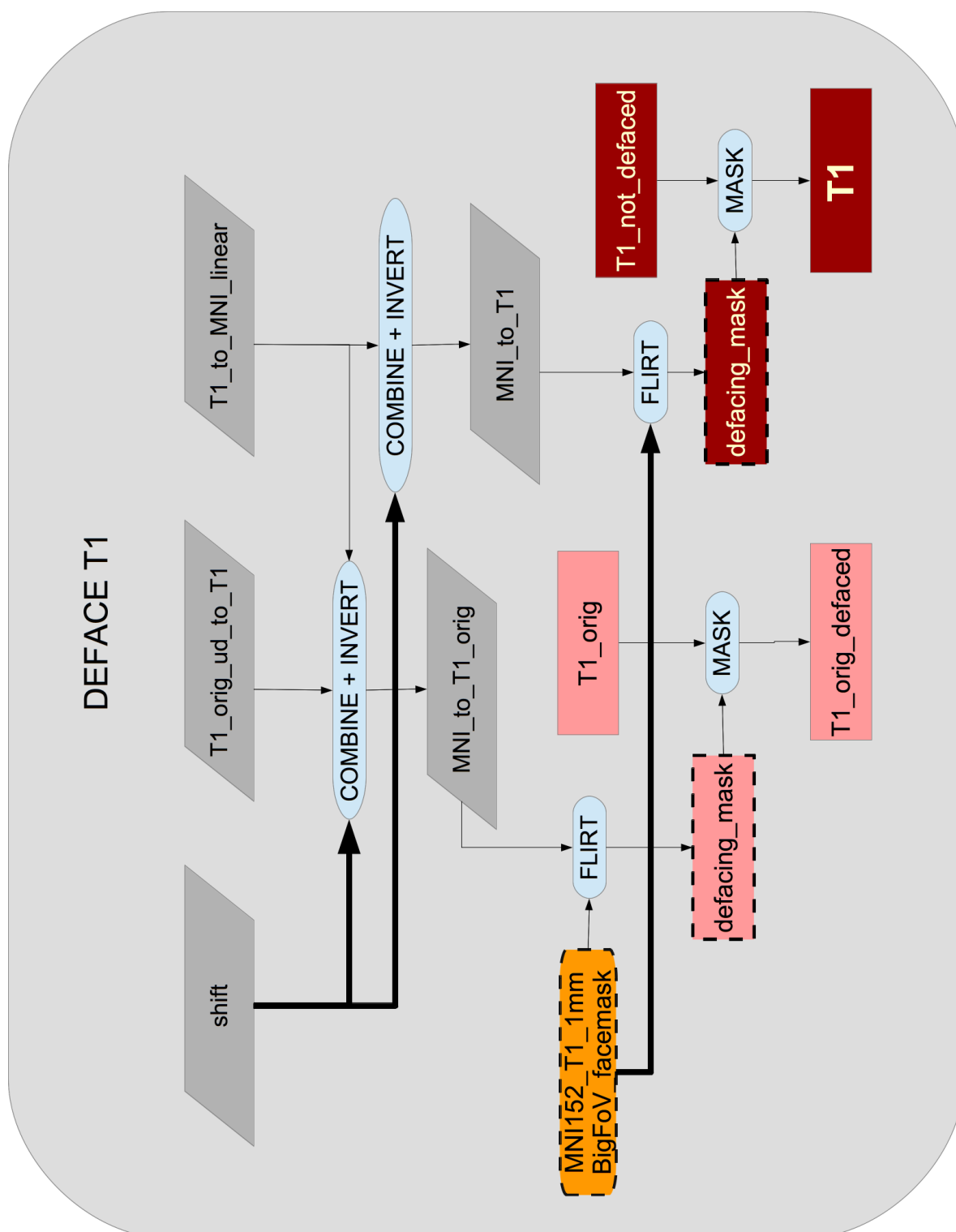


Figure 2.7: Flowchart for the defacing of the T1.

Next, tissue-type segmentation is applied using FAST (FMRIB’s Automated Segmentation Tool [Zhang et al. \(2001\)](#)). This estimates discrete and probabilistic segmentations for CSF (cerebrospinal fluid), grey matter and white matter. As part of the segmentation, intensity bias is estimated, and so this step is also used to generate a fully bias-field-corrected version of the brain-extracted T1.

These data are then used to carry out a SIENAX¹⁵ analysis [Smith et al. \(2002\)](#). The external surface of the skull is estimated from the T1 and used to normalise brain tissue volumes for head size (compared with the MNI152 template). Volumes of different tissue types and total brain volume, both unnormalised and normalised for head size, are then generated.

Subcortical structures (shapes and volumes) are modelled using FIRST (FMRIB’s Integrated Registration and Segmentation Tool [Patenaude et al. \(2011\)](#)). The shape and volume outputs for 15 subcortical structures are generated and stored. A single summary image, with a different value coding for each structure, is also generated.

From the T1 structural image, several global volume measures from SIENAX are reported as distinct IDPs, both normalised for overall head size as well as not normalised: total brain (grey + white matter) volume; total white matter volume, total grey matter volume, ventricular (non-peripheral) CSF (cerebrospinal fluid) volume; peripheral cortical grey matter volume. The overall volumetric head-size scaling factor is also recorded as an IDP. Several subcortical structures’ volumes from FIRST segmentation (not normalised for brain/head size, although that normalisation can be easily done, as the total brain volume is an IDP) are also reported, in general with separate IDPs for left and right, such as left thalamus and right thalamus. The total volume of grey matter (using the grey matter partial volume estimates from FAST) within 139 GM ROIs¹⁶ and total volume of white matter hyperintensities (WMHs) calculated with BIANCA ([Griffanti et al., 2016](#)) (using both T1 and T2 FLAIR) are also included as IDPs.

2.2.3 T2 FLAIR pipeline

The T2 FLAIR processing pipeline is very similar to the T1 pipeline, although we use the non-linear T1-to-MNI152 transformation to transform T2 FLAIR to MNI152 space.

¹⁵Structural Image Evaluation using Normalisation of Atrophy: Cross-sectional

¹⁶The ROIs are defined by the combination of parcellations from several atlases: Harvard-Oxford cortical and subcortical atlases, and Diedrichsen cerebellar atlas

For this reason and to assist other combined analyses (e.g., white-matter hyperintensity segmentation), the T2 FLAIR image is first transformed to T1 space as described below.

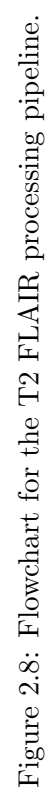
As shown in Figure 2.8, the original T2 FLAIR image is first corrected for gradient distortions (GDC), and then a rigid-body (6 degrees of freedom) linear registration using FLIRT is applied to transform the corrected T2 FLAIR into corrected T1 space. After this step, the T1 brain and defacing masks are applied to the T2 FLAIR image (see Figure 2.9). Finally, an MNI152-version of T2 FLAIR is also generated using the previously calculated warp from T1 to MNI152.

As the last step, we use the estimated bias field previously calculated by FAST on the T1 image to correct the residual bias field inhomogeneities in the T2 FLAIR image. This approach performs as well as estimating the bias field directly from the T2 FLAIR that has not been removed by the on-scanner “pre-scan normalise” processing. To validate this approach, we confirmed that the histograms for white matter intensity in 78 (randomly chosen) subjects improved, i.e., the distribution of WM intensity (as measured by the Inter Quartile Range [IQR]) was tightened similarly for T1 and T2 FLAIR after using on-scanner bias-field correction and T1’s FAST bias field correction. Figure 2.10 (panels A and B) shows the intensities for a typical subject, with a similar improvement for T1 and T2 FLAIR evident. Figure 2.10.C shows the different inter-quartile-ranges for all 78 subjects, with the results further summarised in Table 2.1.

2.2.4 swMRI pipeline

The processing pipeline for this modality is shown in Figure 2.11.

Combining phase images across coils requires care due to anomalous phase transitions in regions of focal signal dropout for a given coil. Currently, all coil channels are saved separately to enable the combination of phase images in post-processing. Each coil channel phase image is first high-pass filtered to remove low-frequency phase variations (including both coil phase profiles and field distortion from bulk shape). A combined complex image is generated as the sum of the complex data from each coil (unfiltered magnitude and filtered phase), and the final phase image is the phase of the summation of the second echo since it has greater venous contrast. Careful inspection of a small number of subjects found no anomalous phase transitions from individual channels in the final combined image.



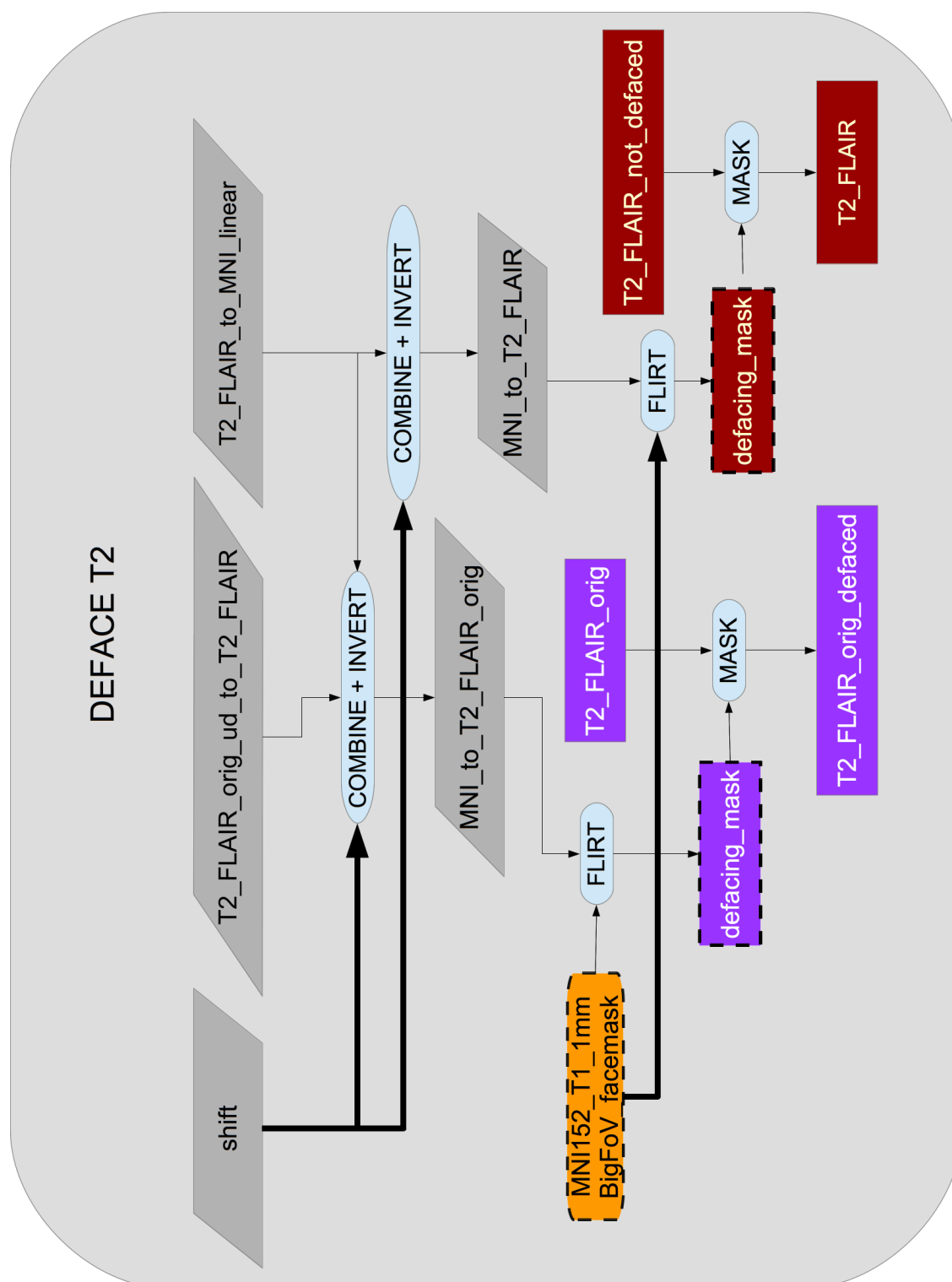


Figure 2.9: Flowchart for the defacing of the T2 FLAIR

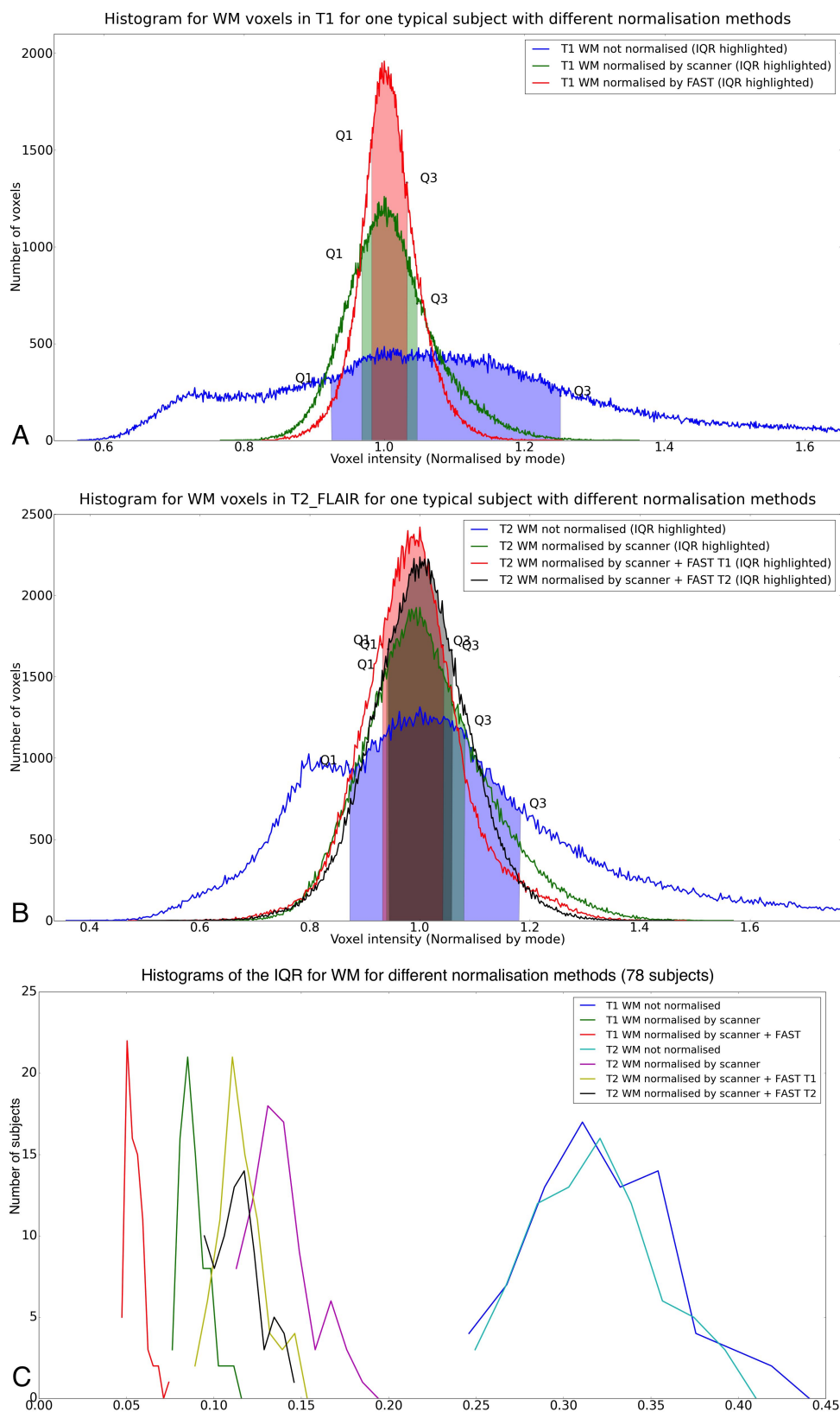


Figure 2.10: Histogram of WM intensities for a typical subject, showing that use of the bias field estimated from the T1 image can be applied to the T2 FLAIR image with similar improvement in IQR range to the one obtained in T1. Also, the histogram of this method is not very different from the one obtained by applying FAST directly to T2 FLAIR. *A*: T1. *B*: T2 FLAIR. *C*: Distribution of WM intensity Inter Quartile Ranges (IQRs) using different normalisation methods in 78 subjects.

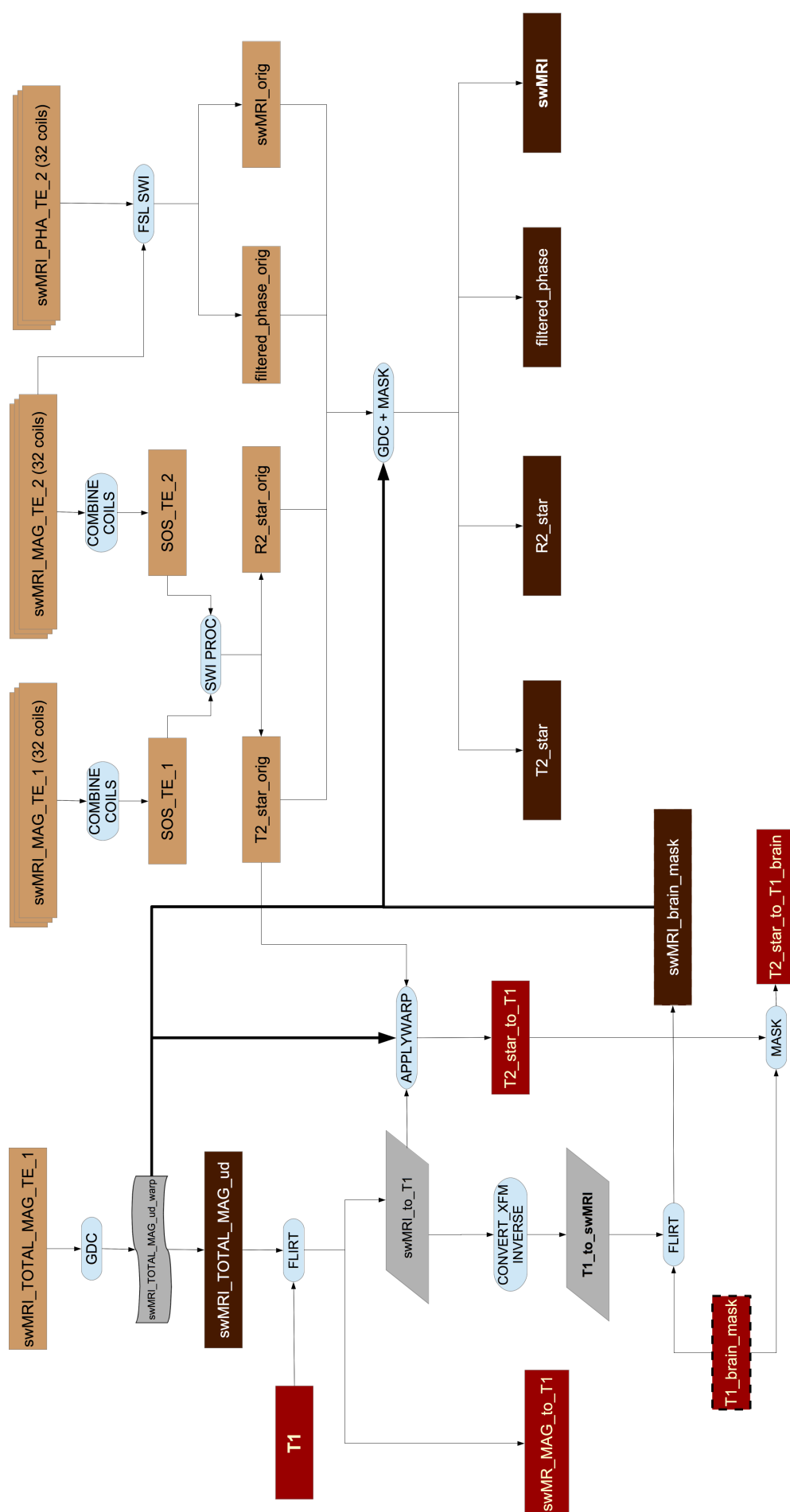


Figure 2.11: Flowchart for the swMRI processing pipeline.

Table 2.1: Descriptive statistics of WM intensity Inter Quartile Ranges (IQRs) using 3 normalisation methods (78 subjects).

Normalisation method	IQR Typical Subject	Mean IQR	Median IQR	Std IQR
T1 not normalised	0.3274	0.3232	0.3191	0.0418
T1 normalised by scanner	0.0797	0.0893	0.0872	0.0086
T1 normalised by scanner + FAST	0.0508	0.0550	0.0535	0.0058
T2 FLAIR not normalised	0.3106	0.3181	0.3163	0.0365
T2 FLAIR normalised by scanner	0.1429	0.1386	0.1356	0.0181
T2 FLAIR normalised by scanner + FAST from T1	0.1116	0.1162	0.1138	0.0136
T2 FLAIR normalised by scanner + FAST from T2 FLAIR	0.1154	0.1144	0.1123	0.0137

Venograms were calculated using an established reconstruction ([Haacke et al., 2004](#)), in which magnitude images are multiplied by further filtering of the phase data to enhance the appearance of veins. The filtering consists of thresholding the phase image to set the phase to zero in voxels indicating diamagnetic susceptibility, and to take the phase to the fourth power in voxels indicating paramagnetic susceptibility. The chosen power represents a trade-off between venous-tissue contrast and noise in the phase data. This enhances image contrast around veins, which are the source of the strongest paramagnetic susceptibility.

Additionally, $T2^*$ values are calculated from the magnitude data. First, the inverse of $T2^*$, termed $R2^*$, is calculated from the magnitude images from the two echo times ($TE1$ and $TE2$). The log of the ratio of these two echo time images is calculated and scaled by the echo time difference, to give the $R2^*$.

$T2^*$ is calculated as the inverse of $R2^*$. The $T2^*$ image is then spatially filtered to reduce noise (3x3x1 median filtering followed by limited dilation to fill small holes of missing data) and transformed into the space of the T1 (via linear registration of the bias-field-normalised first-echo magnitude image). The median (across ROI voxels) $T2^*$ value is then estimated as a separate IDP for each of the subcortical structure ROIs (left thalamus,

right caudate, etc.) obtained from the T1.

2.2.5 Fieldmap generation pipeline

Fieldmap images reflect variations in the static magnetic field. These are needed to correct geometric distortions in the phase encoding direction in EPI images (Jezzard and Balaban, 1995).

In the fieldmap generation pipeline (see Figure 2.12) we estimate the fieldmaps from the $b=0$ images¹⁷ in the dMRI data using the opposing AP - PA phase-encoding acquisitions mentioned above. The reasons for doing so, rather than measuring the field using a dual echo-time gradient echo acquisition, are numerous. Firstly, “traditional” fieldmaps (based on dual echo-time gradient-echo images) are not free of problems (Andersson and Skare, 2010). The choice of echo-time difference represents a trade-off between wanting a large echo-time difference to get a large effect to calculate the field from, and wanting a small echo-time difference which results in fewer phase-wraps that may be hard to unwrap. Furthermore, such a fieldmap will not automatically be in the same space (position) as the images one wants to correct. Therefore the fieldmap needs to be registered to those images, and that is not a trivial task as they are differently distorted. Also, as mentioned in Smith et al. (2013), fieldmaps estimated with reversed encoding direction acquisitions produce an “equivalently good distortion correction accuracy” to that achieved with traditional fieldmaps, while being faster to acquire. This reduction in time, aside from making the acquisition less susceptible to within-scan head motion, saves acquisition time, which is a valuable resource in UK Biobank.

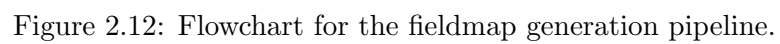
All $b=0$ dMRI images (interspersed with the high- b images) with opposite phase-encoding direction (AP and PA) are analysed to identify the most suitable pair of AP, and PA images¹⁸. This is achieved by aligning all AP images to each other with a rigid-body registration (6 degrees of freedom) and then calculating the correlation of each $b=0$ image with all others. The volume that best correlates with all others is selected¹⁹ as the “target”, subject to the following additional criterion.

As the chosen $b=0$ image determines the space for later processing, we want that space to be as close to the space of the first image with diffusion encoding ($b>0$) as possible.

¹⁷ $b=0$ images (also called b_0) have a b -value below 50

¹⁸Differences between $b=0$ images in each encoding direction can occur due to subject movement

¹⁹All steps applied to images in the AP direction are also applied to images in the PA direction



The movement estimation between the $b=0$ image (no diffusion weighting) images and the two shells with diffusion encoding is not carried out until the end of the eddy current correction, so if the fieldmap is far away from that first dMRI, the movement estimation process will be less optimal. Furthermore, dMRI images are acquired shortly after the fMRI images, so the case for selecting an early $b=0$ image is stronger²⁰. Hence, we apply a selection bias towards the first $b=0$ image: if the first $b=0$ image has sufficient quality (correlation of 0.98 or higher to the other images), we select it as the “best $b=0$ image” (the same selection bias is applied to PA).

This optimal AP/PA pair is then fed into topup²¹ (Andersson et al. (2003)) in order to estimate the fieldmap and associated dMRI EPI distortions.

The full generation method of the fieldmap magnitude is explained in Section 2.2.6. GDC is applied to this magnitude image before registering it to the T1 (which has already been gradient-distortion corrected). The result is linearly aligned to the T1 (using boundary-based-registration [BBR] as the cost function as described in Greve and Fischl (2009)) for later use in unwarping the fMRI data. The resulting transformation is applied to the fieldmap and also inverted to get the dilated structural brain mask in the fieldmap (dMRI) space. Finally, the fieldmap image in structural space is brain-masked and converted to radians per second for later use for fMRI unwarping.

2.2.6 Fieldmap magnitude generation

Running topup returns several outputs like the phase of the estimated fieldmap, the off-resonance field used to correct the geometric distortions, and the corrected input $b=0$ pair of images. Usually, one of those $b=0$ images (or the average) is taken as the magnitude fieldmap image (used for registering the dMRI data to structural space), but that may cause some problems in our data due to the tight FOV. These magnitude images will only have non-zero values for voxels where data exists for both images of the pair. Therefore, in the presence of subject movement between the images, there can be zero-voxels introduced into the data.

²⁰The fieldmaps will be used to correct both fMRI and dMRI data

²¹After making sure that the number of slices in the Z direction is a multiple of topup’s sub-sampling level (See <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/topup/TopupUsersGuide#A--subsamp>), defined in topup configuration file, by removing the appropriate number of slices from the top of the image

Therefore, to get a correct fieldmap magnitude, the following steps are needed in both AP and PA (see Figure 2.13):

- Apply topup warp to the original $b=0$ image.
- Generate a mask with the zero-valued voxels from the corrected $b=0$ image from the previous step.
- Multiply each of those masks by the corrected $b=0$ image of the opposite direction and add the result to the corrected $b=0$ image. That way, for each zero-valued voxel in the corrected $b=0$ image, we get the corresponding value from the corrected opposite $b=0$ image.
- Average these.

2.2.7 dMRI pipeline

As can be seen in Figure 2.14, in the first step of this part of the pipeline the dMRI data (AP encoding direction) is corrected for eddy currents and head motion and has outlier-slices (individual slices in the 4D data) corrected using the eddy tool (Andersson and Sotiropoulos, 2015, 2016; Andersson et al., 2016). This step requires knowing the “best” $b=0$ image in the AP direction as discussed above. The primary corrections carried out by eddy need to be done in-plane and applying the GDC before eddy would move data out of the plane. Therefore, GDC is applied after eddy to produce a more accurate correction, as explained in Glasser et al. (2013).

The output is fed into two complementary analysis, one based on tract-skeleton (TBSS) processing, and the other based on probabilistic tractography (bedpostx / probtrackx). Both analysis streams then report a range of dMRI-derived measures within different tract regions: A) measures derived from diffusion-tensor modelling, and B) measures derived from microstructural model fitting.

2.2.7.1 Diffusion-Tensor-Image fitting and NODDI

The $b=1000$ shell (50 directions) is fed into the DTI fitting (Basser et al., 1994) tool DTIFIT, creating outputs such as fractional anisotropy, tensor mode and mean diffusivity.

In addition to the DTI fitting, the full two-shell dMRI data is fed into NODDI (Neurite Orientation Dispersion and Density Imaging) (Zhang et al., 2012) modelling, using the AMICO (Accelerated Microstructure Imaging via Convex Optimization) tool (Daducci

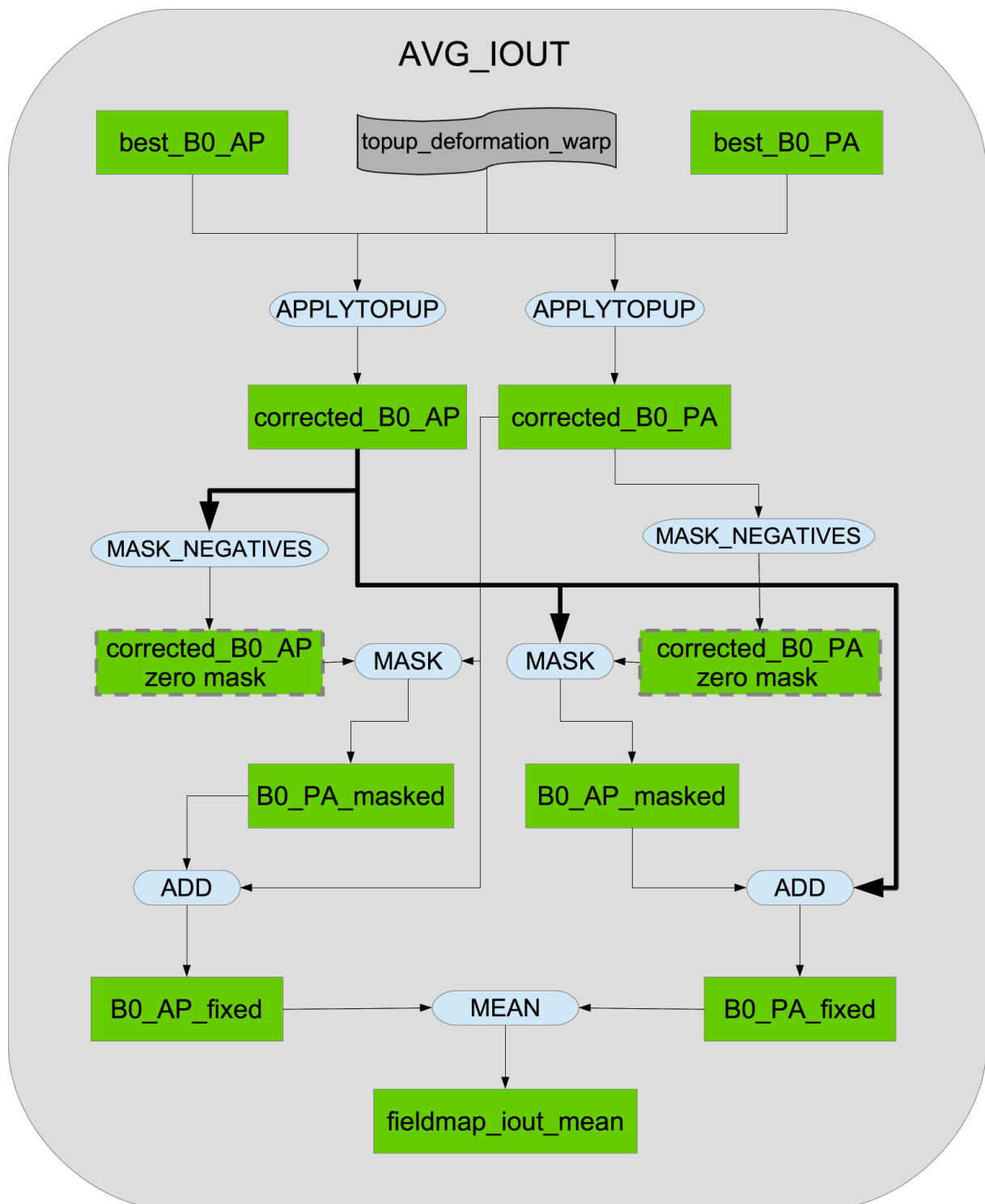
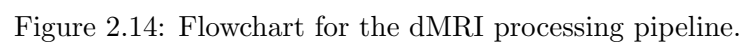


Figure 2.13: Flowchart for the averaging of the B0s for the fieldmap generation.



et al., 2015). This aims to generate meaningful voxelwise microstructural parameters, including ICVF (intra-cellular volume fraction - an index of white matter neurite density), ISOVF (isotropic or free water volume fraction) and ODI (orientation dispersion index, a measure of within-voxel tract disorganisation).

2.2.7.2 WM tract skeleton analysis

The DTI Fractional Anisotropy (FA) image is fed into TBSS (Smith et al., 2006), which aligns the FA image onto a standard-space white-matter skeleton, with alignment improved over the original TBSS skeleton-projection methodology through the utilisation of a high-dimensional FNIRT-based warping (de Groot et al., 2013). This decision was re-validated along the lines of de Groot et al. (2013) after a thorough comparison of 14 different alignment methods applied to UK Biobank data (see Figure 2.15). The resulting standard-space warp is applied to all other DTI/NODDI output maps. For each of the DTI/NODDI maps, the skeletonised images are averaged within a set of 48 standard-space tract masks defined by the group of Susumu Mori at Johns Hopkins University Mori et al. (2005) and Wakana et al. (2007), (similar to the TBSS processing applied in the ENIGMA project Jahanshad et al. (2013)), to generate a set of dMRI IDPs.

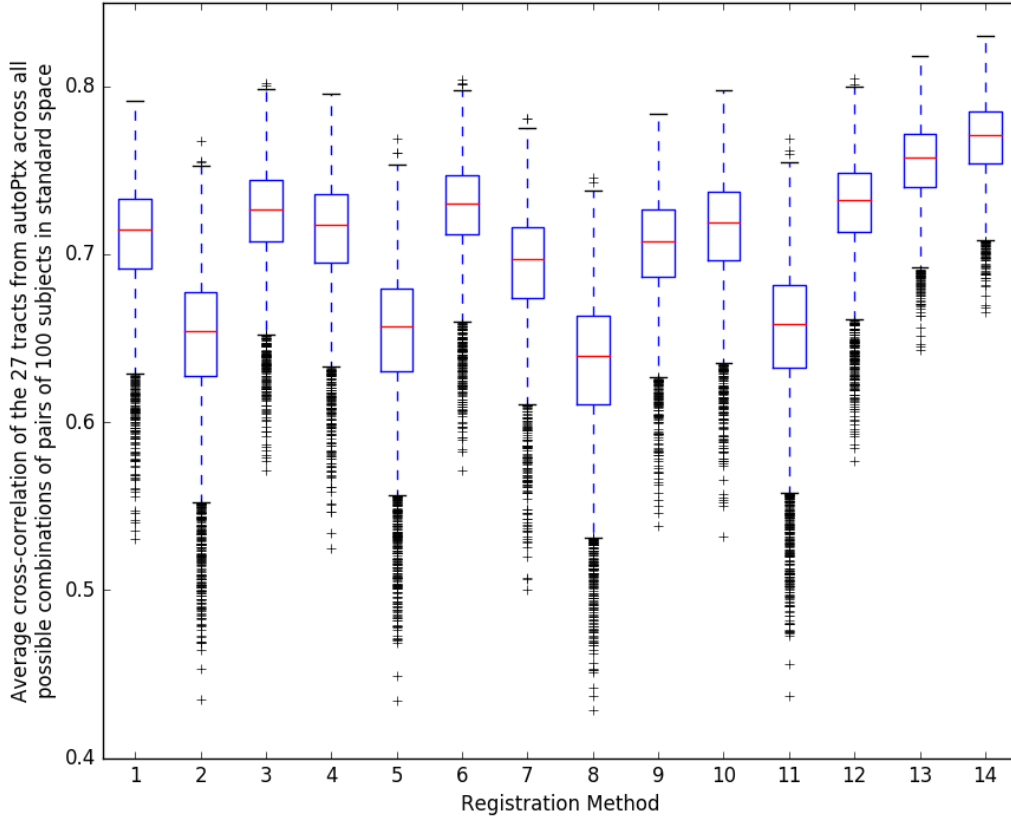


Figure 2.15: Comparison of 14 different alignment methods of FA to MNI space. We used the same methodology as [de Groot et al. \(2013\)](#). For each registration method, we used its estimated warp field in autoPtx to transform 27 automatically defined tracts into standard space; as discussed in [de Groot et al. \(2013\)](#), judging cross-subject alignment through similarity of tracts can be considered a test of alignment success that is reasonably independent of the images and cost functions used to drive the alignments. Each box plot shows distribution of tract cross-correlation (averaged over the 27 tracts for each pair of subjects) across the 4950 possible pairings of 100 subjects.

- 1: FA linearly aligned to T1 + T1 non-linearly aligned to MNI.
- 2: FA linearly aligned to T1 + T1's WM non-linearly aligned to MNI's WM.
- 3: FA linearly aligned to T1 + T1's GM non-linearly aligned to MNI's GM.
- 4: Corrected b=0 linearly aligned (BBR) to T1 + T1 non-linearly aligned to MNI.
- 5: Corrected b=0 linearly aligned (BBR) to T1 + T1's WM non-linearly aligned to MNI's WM.
- 6: Corrected b=0 linearly aligned (BBR) to T1 + T1's GM non-linearly aligned to MNI's GM.
- 7: FA non-linearly aligned to T1 + T1 non-linearly aligned to MNI.
- 8: FA non-linearly aligned to T1 + T1's WM non-linearly aligned to MNI's WM.
- 9: FA non-linearly aligned to T1 + T1's GM non-linearly aligned to MNI's GM.
- 10: FA linearly aligned (BBR) to T1 + T1 non-linearly aligned to MNI.
- 11: FA linearly aligned (BBR) to T1 + T1's WM non-linearly aligned to MNI's WM.
- 12: FA linearly aligned (BBR) to T1 + T1's GM non-linearly aligned to MNI's GM.
- 13: FA non-linearly aligned to FA FMRIB58 atlas via an FA study-specific template (created by aligning all the FAs to FA FMRIB58 and then averaging).
- 14: FA non-linearly aligned to FA FMRIB58 atlas using high-dimensional FNIRT-based warping.

Figures 2.16, 2.17, 2.18, and 2.19 show the same comparison at the tract level (27 tracts).

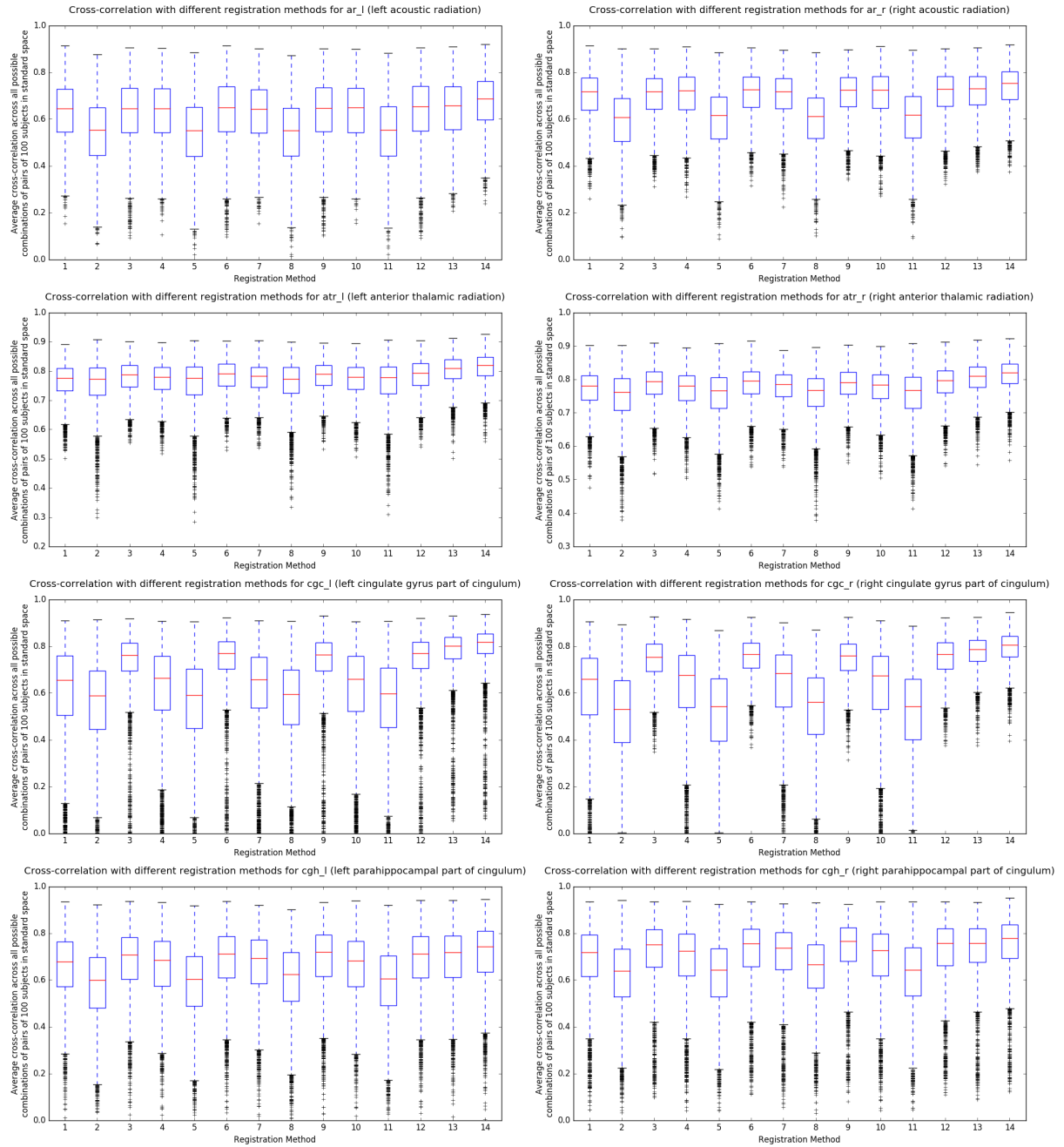


Figure 2.16: Comparison of 14 different alignment methods of FA to MNI for tracts 1 to 8

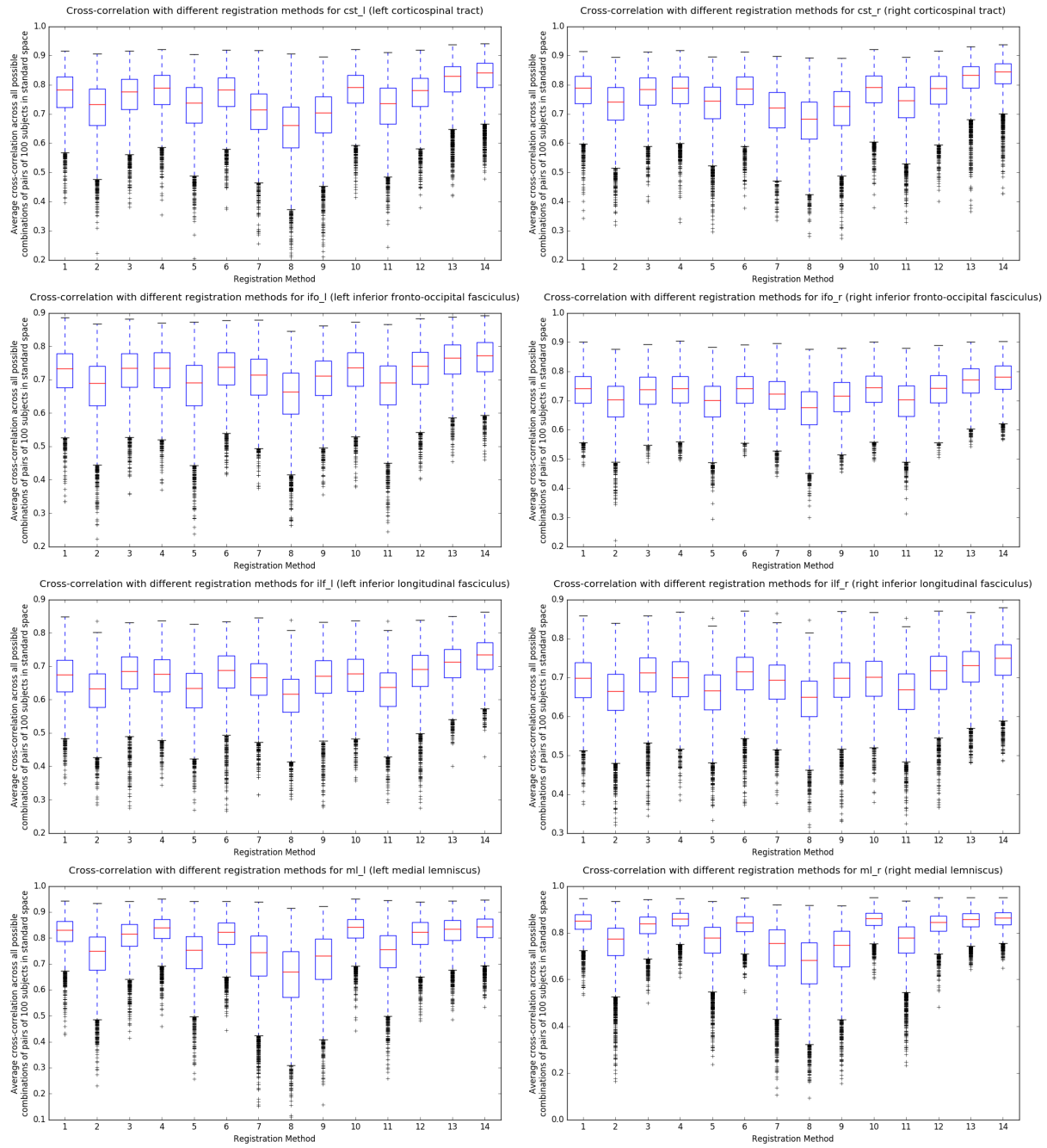


Figure 2.17: Comparison of 14 different alignment methods of FA to MNI for tracts 9 to 16

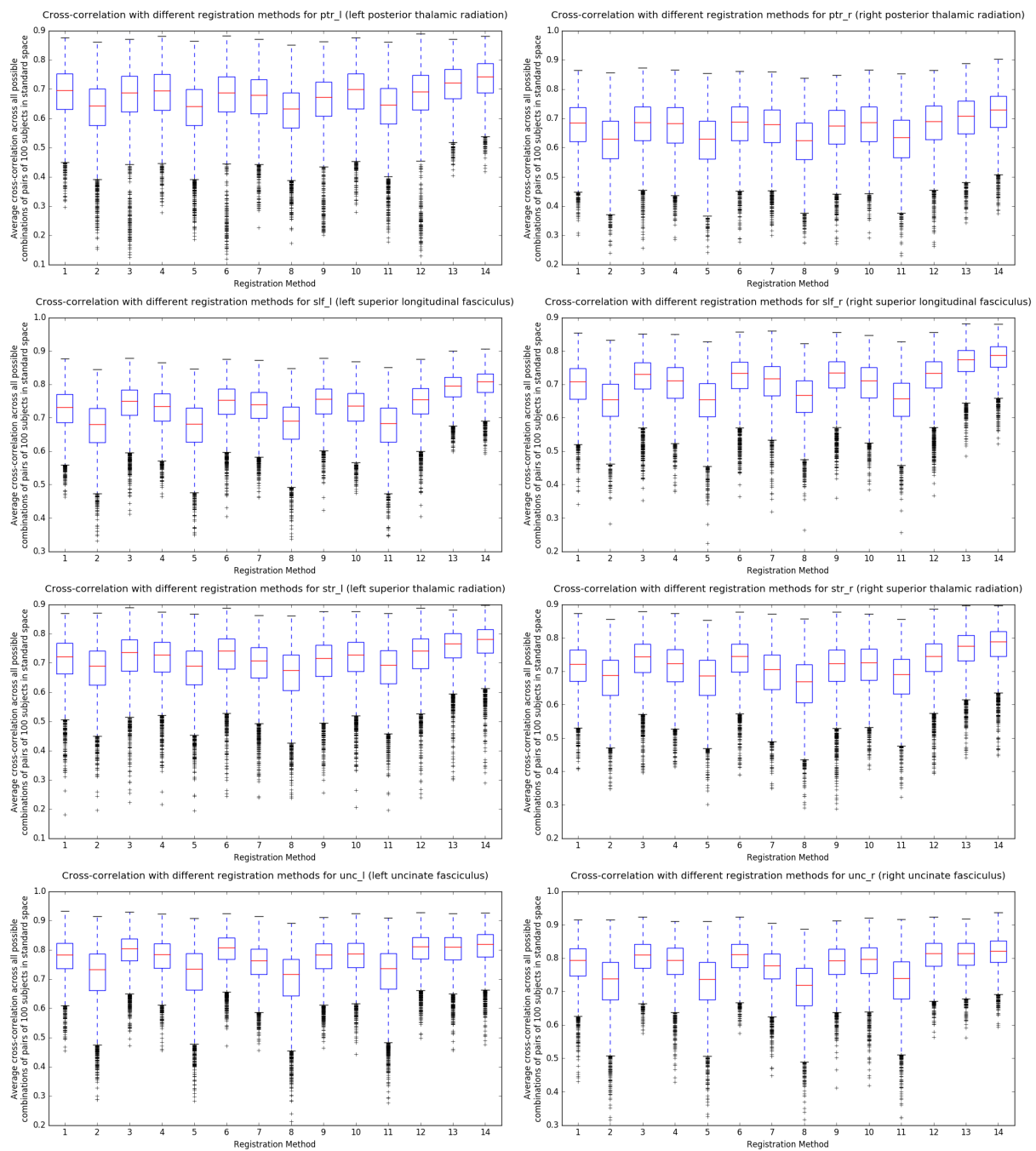


Figure 2.18: Comparison of 14 different alignment methods of FA to MNI for tracts 17 to 24

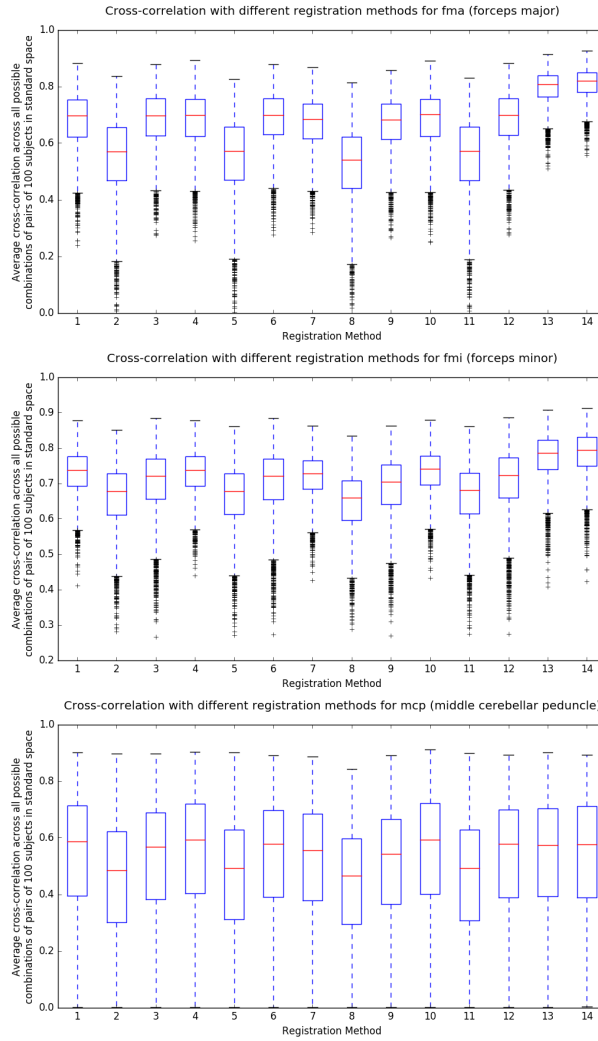


Figure 2.19: Comparison of 14 different alignment methods of FA to MNI for tracts 24 to 27

2.2.7.3 Probabilistic-tractography-based analysis

In addition to the TBSS analyses, the preprocessed dMRI data is also fed into a tractography-based analysis. This begins with within-voxel modelling of multi-fibre tract orientation structure via the bedpostx tool (Bayesian Estimation of Diffusion Parameters Obtained using Sampling Techniques)²², which implements a model-based spherical deconvolution and estimates up to 3 fibre orientations per voxel. This is followed by probabilistic tractography (with crossing fibre modelling) using probtrackx (Behrens et al., 2003, 2007; Jbabdi et al., 2012; Hernández et al., 2013). The bedpostx outputs are suitable for running tractography from any (voxel or region) seeding; the pipeline has already automatically mapped a set of 27 major tracts using standard-space start/stop ROI masks defined by AutoPtx de Groot et al. (2013). In order to reduce the amount of processing

²²Model 2: Deconvolution with sticks and a range of diffusivities

time for those tracts, AutoPtx was modified to reduce the number of seeds per voxel (the relationship between the number of seeds per voxel and processing time is linear) to 0.3 times the number of seeds specified in the original version of AutoPtx. Figure 2.21 illustrates the process we used to select this factor.

Although eddy and bedpostx outputs are in the space and resolution of the (GDC-unwarped) native diffusion data, the non-linear transformation between this space and 1mm MNI standard space (as estimated by TBSS above) is used to create tractography results in 1mm standard space. For each tract and each DTI/NODDI output image type, we compute the weighted-mean value of the DTI/NODDI parameter within the tract (the weighting being determined by the tractography probabilistic output) to generate a different set of dMRI IDPs.

2.2.8 Resting fMRI pipeline

This section of the pipeline uses several outputs from the T1 pipeline (T1, brain-extracted T1, linear and non-linear warp of T1 to MNI152 space and binary mask of the white matter in T1 space). See Figure 2.22 for the flowchart.

Using this information and a previously calculated GDC warp for the rfMRI data, a configuration file is generated in a format suitable for use by Melodic (Beckmann and Smith, 2004). Melodic is a complete pipeline in itself which here performs EPI unwarping (utilising the fieldmaps as described above), GDC unwarping, motion correction using MC-FLIRT (Jenkinson et al., 2002), grand-mean intensity normalisation of the entire 4D dataset by a single multiplicative factor, and high-pass temporal filtering (Gaussian-weighted least-squares straight-line fitting, with $\sigma = 50.0s$). To reduce interpolation artefacts, the EPI unwarping, GDC transformation, and motion correction are combined and applied simultaneously to the functional data. The combination of these warps can be seen in Figure 2.23.

Finally, structured artefacts are removed by ICA + FIX processing (independent component analysis followed by FMRIB’s ICA-based X-noiseifier (Beckmann and Smith, 2004; Salimi-Khorshidi et al., 2014; Griffanti et al., 2014). FIX was hand-trained on 40 Biobank rfMRI datasets following the methodology described in Griffanti et al. (2017), and leave-one-out testing showed (mean/median values across subjects) 99.1/100.0% classification accuracy for non-artefact components and 98.1/98.3% accuracy for artefact components.

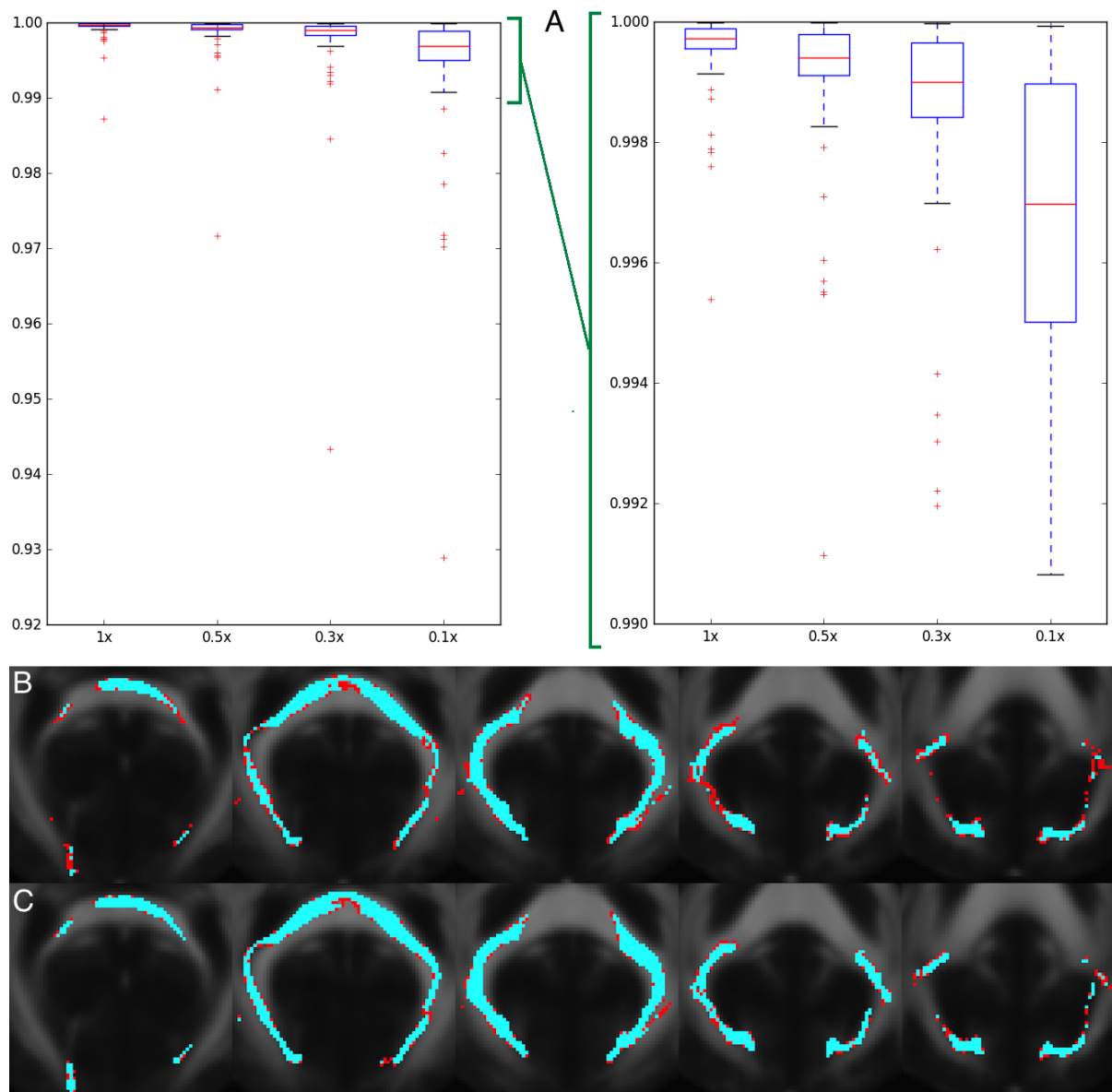


Figure 2.21: Degree of correlation and similarity of tracts after reducing the number of seeds per voxel in probtrackx. *A*: Average over 27 tracts and five subjects of the correlation between 2 different runs of probtrackx; this is reduced when we reduce the number of seeds per voxel. The X-axis is the reduction in the number of seeds for the original AutoPtx configuration. The Y-axis is the average correlation of the pairs of tracts. When we reach 0.1x seeds per voxel, the median correlation drops below 0.999 for some tracts. The right plot is a zoom of the left plot. *B*: Worst instance (in terms of correlation) of probtrackx in the forceps major tract with a factor of 0.1x seeds per voxel. The results remain robust, even using a reduced number of seeds per voxel. Correlation between the maps for these two runs was **0.929**. Tracts were binarised for visualisation using a threshold of 10% of the 99th percentile. The overlap of the two runs is shown in blue. The difference is shown in red. FMRIB58 FA atlas is shown in the background. *C*: Same tract (forceps major) from the same subject with a factor of 0.3x seeds per voxel. The results improve by increasing the number of seeds per voxel. Correlation between the maps for these two runs was **0.985**. FMRIB58 FA atlas is shown in the background. **The final decision was to use 0.3x.**

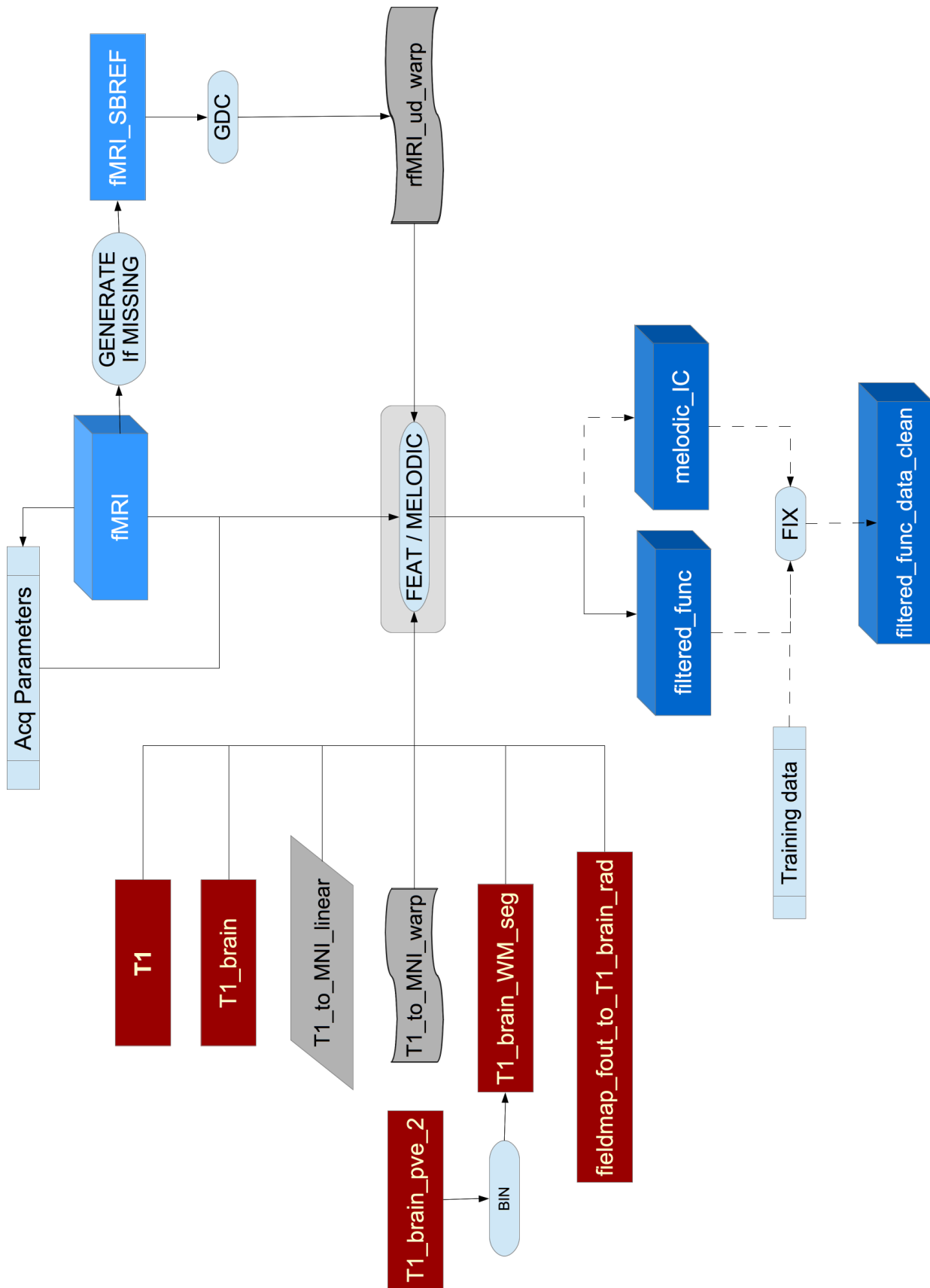


Figure 2.22: Flowchart for the fMRI processing pipeline.

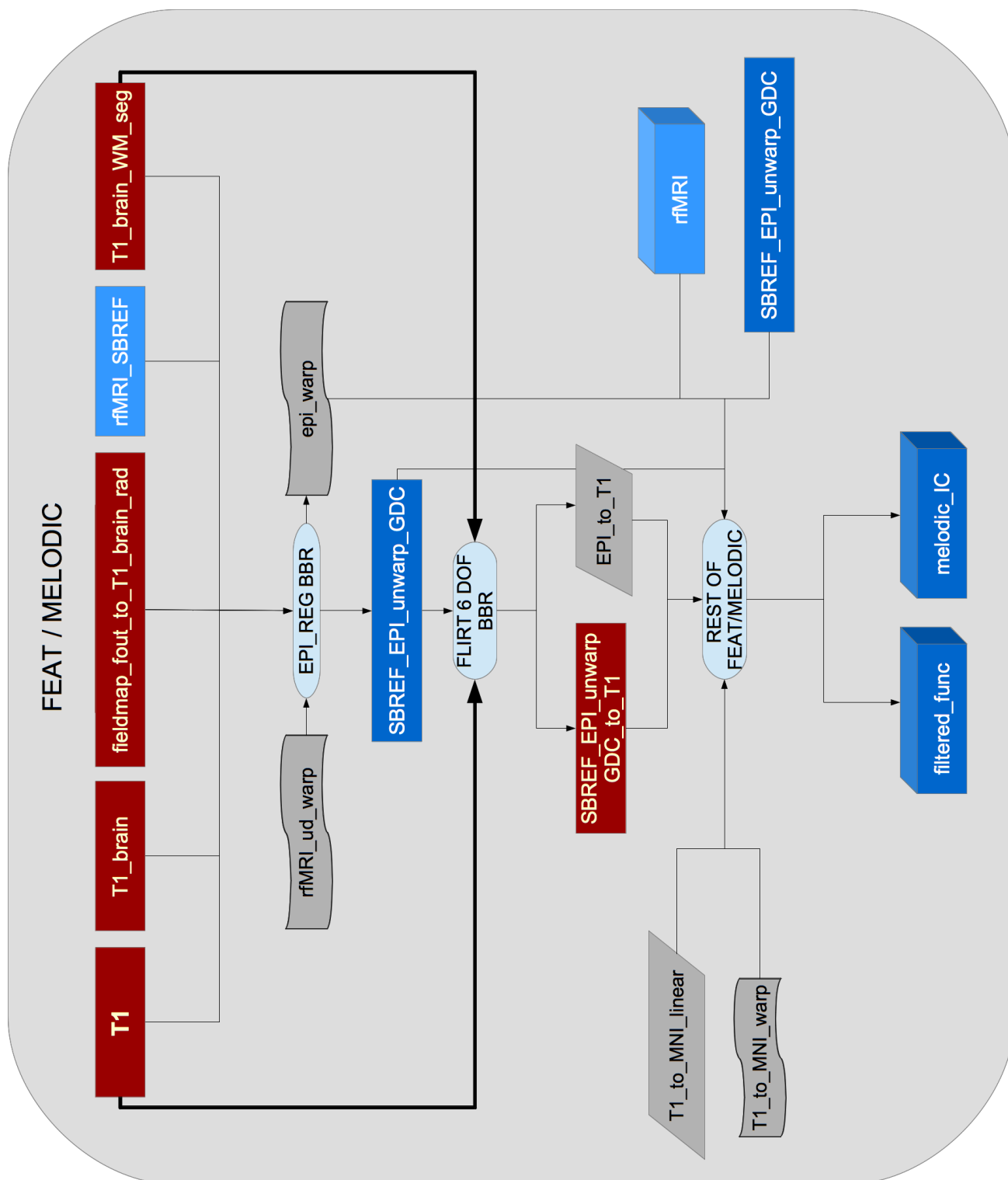


Figure 2.23: Flowchart for the FEAT part of the pipeline.

At this point, no low-pass temporal or spatial filtering has been applied.

We evaluated the effect that FIX had on the relationship between $1/tSNR$ and amount of head motion (summarised to a single average value for each subject). Before FIX cleanup, the correlation was very high ($r=0.75$), indicating strong corruption of data in general by head motion, even after standard (geometric) head motion correction. However, after FIX cleanup, this correlation dropped dramatically ($r=0.1$, i.e., only 1% of variance explained by head motion), indicating high effectiveness of FIX in removing residual motion artefacts.

The EPI unwarping is a combined step that includes GDC and alignment to the T1, though the unwrapped data is stored in native (unwarped) fMRI space (and the transform to T1 space stored separately). This T1 alignment is carried out by FLIRT, using BBR as the cost function. After the fMRI GDC unwarping, a final FLIRT realignment to T1 space is applied, to take into account any shifts resulting from the GDC unwarping. The previously described transform from T1 space to standard MNI space is utilised when fMRI data is needed in standard space.

To be able to identify RSNs (Resting-State Networks) in individual subjects, we first identify a set of RSNs which are common across the entire group. Therefore, group-average RSN analysis was carried out using 4100 datasets.

First, each subject's preprocessed (as above) time-series dataset was resampled into standard space, temporally demeaned and had variance normalisation applied according to Beckmann and Smith (2004). Group-PCA was then carried out by MIGP (MELODIC's Incremental Group-PCA) from all subjects. This comprises the top 1200 weighted spatial eigenvectors from a group-averaged PCA (a very close approximation to concatenating all subjects' time-series and then applying PCA) Smith et al. (2014). The MIGP output was fed into ICA using FSL's MELODIC tool Hyvärinen (1999); Beckmann and Smith (2004), applying spatial-ICA at two different dimensionalities (25 and 100²³). The dimen-

²³The reasoning behind these dimensionalities is:

- 25 results in large scale network decomposition which matches much of the canonical RSN literature (Smith et al., 2009)
- 100 corresponds to a more finely detailed parcellation. We empirically found that it was not useful to go even higher because, with volumetrically aligned data, raising the number of components did not significantly raise the number of non-artefact group-level components (as opposed to using surface-based analysis where the number of well-aligned small components can be much higher (Smith et al., 2013))

sionality determines the number of distinct ICA components; a higher number means that distinct regions within the spatial component maps are smaller. The group-ICA spatial maps are available online in the UK Biobank showcase²⁴. The sets of ICA maps can be considered as “parcellations” of (cortical and sub-cortical) grey matter, though they lack some properties often assumed for parcellations - for example, ICA maps are not binary masks but contain a continuous range of values; they can overlap each other; a given map can include multiple spatially separated peaks/regions. Any group-ICA components that are clearly identifiable as artefactual (i.e., not neuronally driven) are discarded for the network modelling described below. A text file is supplied with the publicly available group-ICA maps, listing the group-ICA components kept in the final network modelling. From the 25-dimensional group-ICA, 21 components were kept as non-artefactual, and from the 100-dimensional group ICA, 55 were kept.

For a given parcellation (group-ICA decomposition of D components), the set of ICA spatial maps was mapped onto each subject’s rfMRI time-series data to derive one representative time-series per ICA component (for these purposes each ICA component is considered as a network “node”). For each subject, these D time-series can then be used in network analyses, described below. This is the first stage in a dual-regression analysis [Filippini et al. \(2009\)](#).

The node time-series are then used to estimate subject-specific network-matrices (also referred to as “netmats” or “parcellated connectomes”). For each subject, the D node-timeseries were fed into network modelling after regressing the time-series of the artefactual nodes out of all others, and then discarding them, leaving D_g nodes. This results in a $D_g \times D_g$ matrix of connectivity estimates. Network modelling was carried out using the FSLNets toolbox²⁵. Network modelling is applied in two ways:

- Using the full normalised temporal correlation between every node time series and every other. This is a common approach and is very simple. However, it has various practical and interpretational disadvantages including an inability to differentiate between directly connected nodes and nodes that are only connected via an intermediate node ([Smith, 2012](#)), as well as being more corrupted (than partial correlation) by residual shared/global artefacts.

²⁴<http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=9028>

²⁵<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets>

- Using the partial temporal correlation between nodes' time-series. This aims to estimate direct connection strengths better than achieved by full correlation. To slightly improve the estimates of partial correlation coefficients, L2 regularisation is applied (setting $\rho=0.5$ in the Ridge Regression `netmats` option in FSLNets).

Netmat values were transformed from Pearson correlation scores (r-values) into z-statistics, including empirical correction for temporal autocorrelation. Group-average netmats are also available online.

As the matrices are symmetric, only values above the diagonal are kept, and unwrapped (taken column-wise) into a single row of $D_g \times (D_g - 1) / 2$ values per subject. This results in one "compound" IDP (containing all network matrix values for a given subject) for each original dimensionality ($D=25$ and 100) and each network matrix estimation method (full correlation and partial correlation).

2.2.9 Task fMRI pipeline

The same preprocessing and registration was applied as for the rfMRI described above, except that spatial smoothing (using a Gaussian kernel of FWHM 5mm) was applied before the intensity normalisation, and no ICA + FIX artefact removal was performed, both decisions being primarily driven by the shorter time-series in the tfMRI (than the rfMRI) and because of the greater general reliance in tfMRI analysis on voxelwise time-series modelling (as opposed to multivariate spatiotemporal analyses common in resting-state fMRI).

Pre-processing and task-induced activation modelling was carried out using FEAT (fMRI Expert Analysis Tool); time-series statistical analysis was carried out using FILM with local autocorrelation correction [Woolrich et al. \(2001\)](#). The timings of the blocks of the two task conditions (shapes and faces) are defined in 2 text files. Five activation contrasts were defined (Shapes, Faces, Shapes+Faces, Shapes-Faces, Faces-Shapes), and an F-contrast also applied across Shapes and Faces.

The three contrasts of most interest are 1 (Shapes), 2 (Faces) and 5 (Faces-Shapes), with the last of those being of particular interest for amygdala activation. Group-average activation maps were derived from analysis across all subjects and used to define ROIs for generating tfMRI IDPs. Four ROIs were derived; 1 (Shapes group-level fixed-effect z-

statistic, thresholded at $Z > 120$ ²⁶); 2 (Faces group-level fixed-effect z-statistic, thresholded at $Z > 120$); 5 (Faces-Shapes group-level fixed-effect z-statistic, thresholded at $Z > 120$); 5a (Faces-Shapes group-level fixed-effect z-statistic, thresholded at $Z > 120$, and further masked by an amygdala-specific mask). The group-average activation maps and ROIs are available online in the UK Biobank showcase mentioned above.

The Featquery tool was used to extract summary statistics for these four contrast/mask combinations, for both activation effect size (expressed as a % signal change relative to the overall-image-mean baseline level) and statistical effect size (z-statistic), with each of these summarised across the relevant ROI in two ways - median across ROI voxels and 90th percentile across ROI voxels.

Display of the task video and logging of participant responses is carried out by ePrime software, which provides several response log files from each subject. These are not used in the above analyses (as the timings of the task blocks are fixed and already known, and the correctness of subject responses are not used in the above analyses), but are available in the UK Biobank database.

Halving task fMRI acquisition time

In order to know if including a new modality was possible (ASL - arterial spin labelling, a semi-quantitative measure of baseline blood flow), we needed to determine how the necessary reduction of acquired timepoints for task fMRI would affect the estimated task-based activations. To that end, we re-analysed the original task fMRI of 100 subjects, keeping only the first half of their timepoints (hence reducing the data from 4 minutes to 2). We performed the same standard task-induced activation modelling with FEAT that was carried out with the original data (reducing the number of task blocks accordingly). We then performed a second-level analysis across the results for these subjects and compared the first- and second-level results with those of the original results. As the Figure 2.24 shows, qualitative results were quite similar; Figure 2.25 shows that group-level statistics were reduced by at most 10%, while Figure 2.26 shows within-subject summary measures (similar to the IDPs) correlated across subjects (for the most important/sensitive contrast, faces-shapes) with $r=0.76$. We concluded that reducing the task fMRI acquisition time without too much loss of accuracy in the tfMRI IDPs was possible.

²⁶Both mixed-effects and fixed-effects generate huge z-statistics with this many subjects, so we are choosing here to simply work with the “average” activation given by using fixed-effects

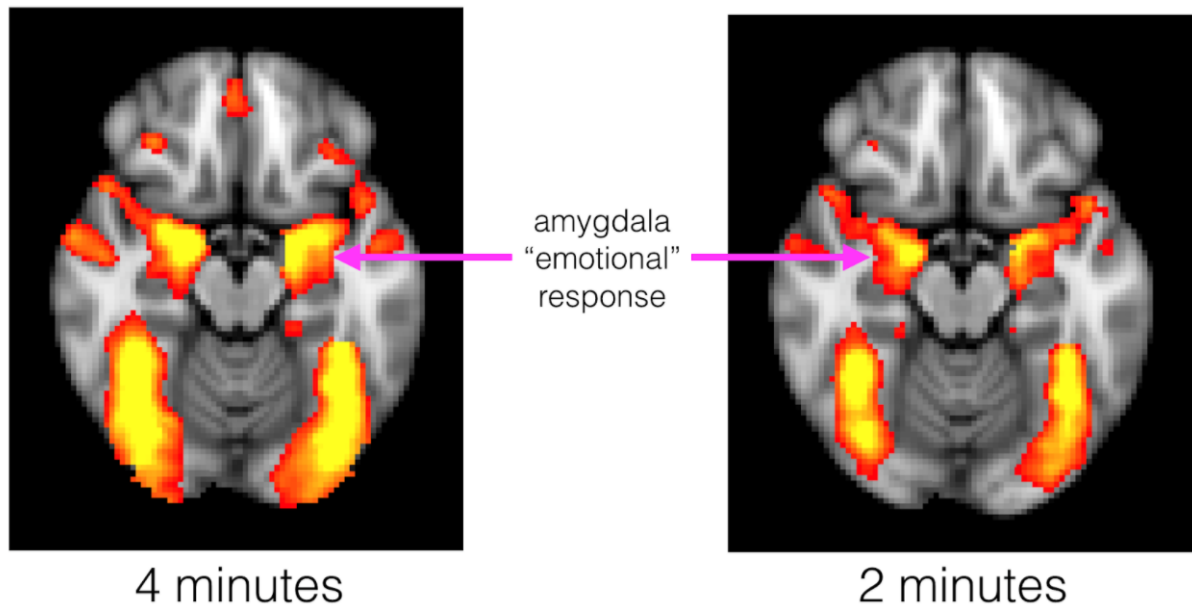


Figure 2.24: Group-level ($n=100$) activation maps for faces-shapes (COPE5) mixed-effects- $Z > 4$

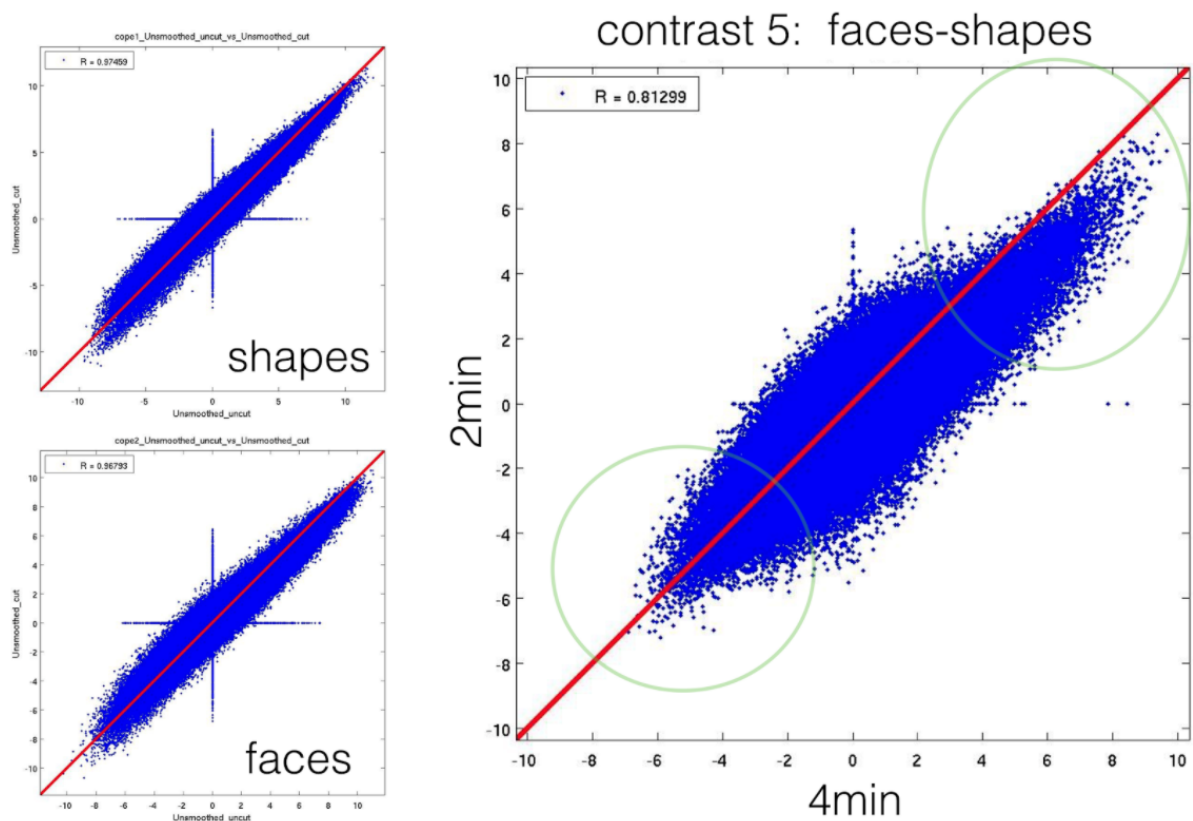


Figure 2.25: Group ME- Z scatter-plots (1 point per voxel) for the three main contrasts: reduced time vs original time. Green circles show how lower activation is similar for both cases, while being reduced by at most 10% in the high Z values.

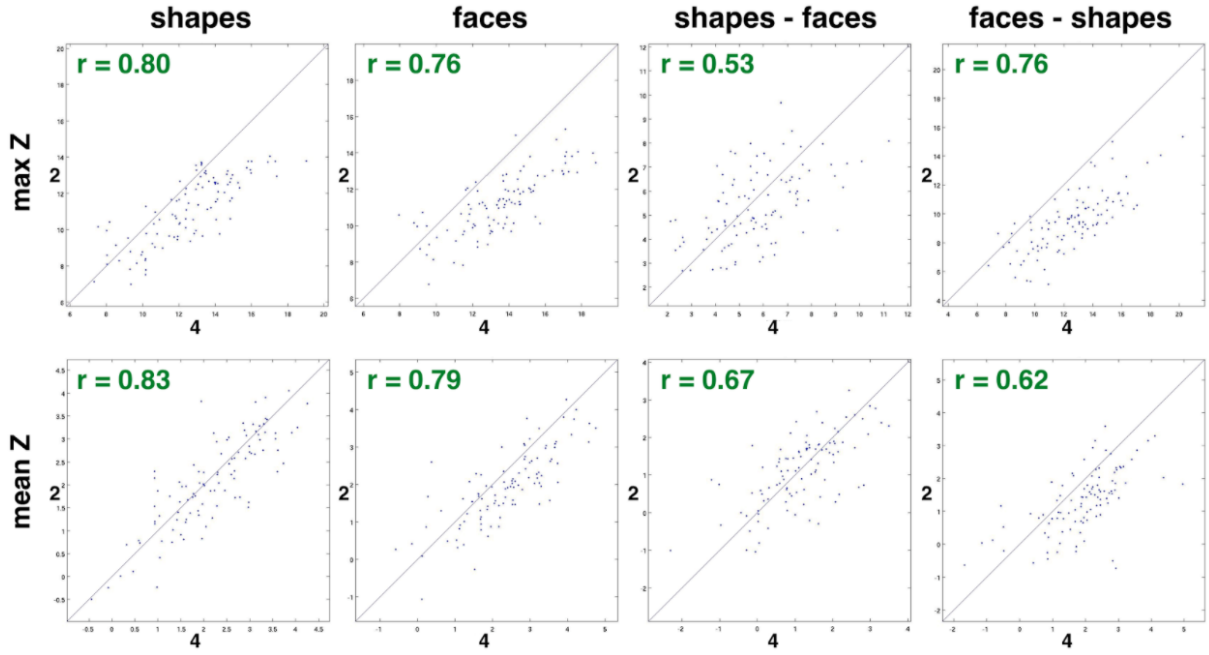


Figure 2.26: Single-subject activation summary results: Peak and mean single-subject Z, inside group-defined activation ROIs (group ME-Z > 5). Each point is a single subject.

2.3 Results: UK Biobank as a resource

When this study ([Alfaro-Almagro et al., 2018](#)) was written, the pipeline had been applied to 10,129 subjects. In November 2015, there was a public release of the imaging datasets and IDPs from the first 5822 subjects ([Miller et al., 2016](#)); a new set of 4282 subjects has recently been processed and publicly released in February 2017 ([Biobank, 2017](#)). The number of datasets from every modality that was deemed suitable for processing is listed in Table 2.2.

Table 2.2: UK Biobank usable modalities in acquisition order. Total subjects: 10,129.

Modality	Available	Usable
T1	10,102	9,933 (98.06%)
rfMRI	10,034	9,558 (94.36%)
tfMRI	9,620	9,182 (90.65%)
T2 FLAIR	9,756	9,045 (89.30%)
dMRI	9,566	8,839 (87.26%)
swMRI	9,396	9,153 (90.36%)
All	8,996	8,211 (81.06%)

Very early improvements in the dMRI and T2 FLAIR protocols were found to be valuable, resulting in large enough data improvements to outweigh the priority of keeping the protocol fixed (and taking into account the relatively small numbers of datasets affected). This change was made at the beginning of protocol.

Expanding on the analyses made in [Miller et al. \(2016\)](#), we have run ~ 8 million univariate

correlations between IDPs and non-brain-imaging variables, to illustrate the statistical power of this resource. This analysis (see Figure 2.27.A) not only shows results largely compatible with the ones presented in Miller et al. (2016) but also that correlations with a P-value smaller than 10^{-100} are possible due to the large number of subjects. Figure 2.27.B demonstrates more directly that the new set of 4282 subjects show very similar characteristics to the first set of 5822 subjects.

Figure 2.28.A shows a straightforward example of the biological relevance of IDPs. In this case, the volume of White Matter Hyperintensities (WMHs) can be used to explore the relationship between age and WMHs. Figure 2.28.B shows an example of the WMH segmentation using BIANCA. This metric also correlates with known health biomarkers, such as systolic and diastolic blood pressure. The positive correlation between total WMH volume and systolic blood pressure (after correcting for the usual confounds²⁷) has a Bonferroni-corrected significance of $P < 10^{-20}$, $r = 0.13$, while the correlation with diastolic blood pressure has a Bonferroni-corrected significance of $P < 10^{-15}$, $r = 0.11$. These significant associations are consistent with the literature (Liao et al., 1996; Gunstad et al., 2005).

2.4 Discussion

The UK Biobank brain imaging data and IDPs can be used to create models that describe the population, and in combination with the healthcare outcomes (which will feed into UK Biobank over coming years), be able to identify risk factors and biomarkers for early diagnosis of many diseases. In this chapter, we provided the first detailed description of the core processing pipeline being developed to process the raw imaging data and generate the IDPs. We also described several investigations necessary to feed into decisions made during the development of the acquisition protocol and the analysis pipeline.

Having a fixed acquisition protocol and a core processing pipeline in UK Biobank is a key feature of this project (Gronenschild et al., 2012; Glatard et al., 2015). Providing a core pipeline that generates both processed data (e.g., with images artefact-cleaned and aligned across modalities and subjects) as well as IDPs will hopefully be valuable to both imaging researchers and non-imaging experts (healthcare researchers, epidemiologists,

²⁷Age, age², sex, age $\times$ sex, age² $\times$ sex, average head motion during tfMRI, average head motion during rfMRI and head size

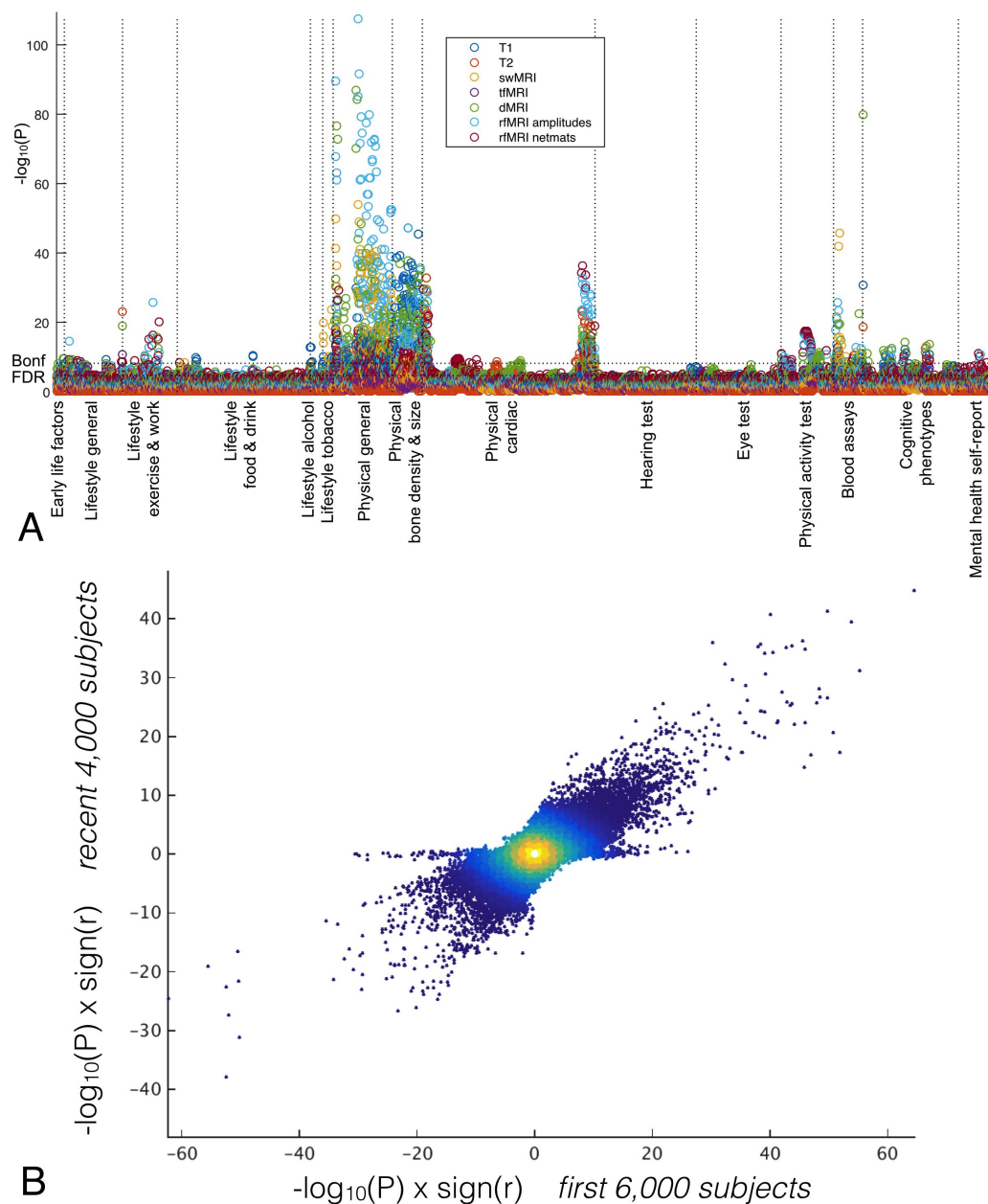


Figure 2.27: *A*: Manhattan plot summarising the significance of 8 million univariate association tests between IDPs and non-brain-imaging variables in the UK Biobank database from 10,000 subjects. For each non-imaging variable (i.e., each column in the plot), only the strongest association is plotted for each class of IDP, for clarity. Plotted p-values are not corrected for multiple comparisons, but the thresholds for both false-discovery-rate and Bonferroni correction are shown as dotted lines. *B*: High reproducibility ($r=0.62$) of these associations in the original vs. new groups of subjects; each point is a given IDP - non-brain-variable pairing. The small number of points along the $y=0$ axis relates to a non-imaging measure which (as a result of investigating these points in this plot) was found to be badly drifting over time).

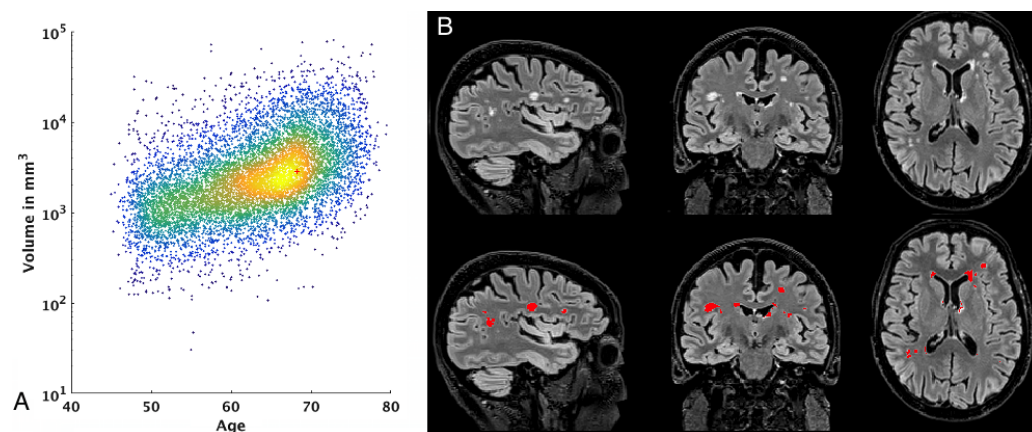


Figure 2.28: *A*: Relationship between age and total volume of white matter hyperintensities. Red cross shows the subject on the right. *B*: WMH segmentation using BIANCA on one example dataset Age: 68.5 years. Total WMH volume: 5049 mm^3 .

etc.). Having said that, the intention is not to discourage other imaging researchers from developing their own image processing and IDP generation approaches and software; indeed, UK Biobank encourages all researchers to feed quantitative research outputs back into the central database. Considering the invaluable connection of the outputs from processing the brain imaging data with the rest of UK Biobank resources and NHS records, the possibilities for future research are massive.

Chapter 3

Quality Control for UK Biobank

3.1 Introduction: The need for Quality Control

Quality Control (QC) is a crucial issue in brain imaging, as a poorly executed application (or lack thereof) can compromise the trustworthiness of a study (Bennett and Miller, 2010). This topic has been explored in the literature, although very often, research is mostly focused on Quality Assurance (QA) rather than QC. QA is focused on avoiding the occurrence of problems by improving a process while QC is focused on finding possible problems in the output of that process. For example, the Function Biomedical Informatics Research Network (FBIRN) set of recommendations (Glover et al., 2012) is solely focused on QA. In a similar vein, Friedman and Glover (2006) explore an interesting set of quality metrics. However, they focus on stability, signal-to-noise ratio (SNR), drift, and other hardware performance issues related to MR scanners, not specifically on the type of artefacts¹ that can be found in MR imaging even when complying with the best QA policies.

The scale of a typical brain imaging study to date (up to 100 subjects) allows researchers to mainly perform QC manually, by visually inspecting the data at each step in the processing pipeline. Even the HCP, while having a relatively large number of subjects (1,200), relies primarily upon visual inspection for their T1 QC process (Marcus et al., 2013). However, the sheer quantity of imaging data which will be produced by UK Biobank (100,000 subjects) makes QC via visual inspection unfeasible. Thus, the development of reliable QC metrics to detect artefacts specific to the imaging acquisition and analysis process is essential.

Various imaging-related QC metrics have been proposed (e.g. Deshmukh and Bhosale (2010), Moorthy and Bovik (2010)) but, as they are not specific to MR or brain imaging, they do not fulfil the needs of our study. More specific proposals can be found in other

¹A comprehensive list of possible artefacts and proposed qualitative quality recommendations can be found in CBS (2016)

studies. Woodard and Carley-Spencer (2006) suggests a metric based on Natural Scene Statistics to detect noise in structural images. Following the same idea, Mortamet et al. (2009) suggests two metrics to detect noise (based on investigating the ratio of artefactual voxels relative to the background, or the noise distribution in the background). Those proposals, although extremely useful, refer to a particular problem in structural images, and hence are unlikely to detect artefacts due to other sources, or in other modalities.

Similarly, there have been proposed metrics for dMRI (Hasan, 2007; Liu et al., 2010; Bastiani et al., 2019), fMRI (Greve et al., 2011; Power et al., 2012; Nichols, 2013; Afyouni and Nichols, 2018), and more recent multimodal approaches (Abe et al., 2015), and even though we have used some of these metrics as confounds (see Chapter 4), the extent to which these metrics may capture the majority of artefacts that could arise is yet to be proven.

Finally, PCP QAP (Preprocessed Connectomes Project Quality Assessment Protocol)² is a concerted effort to create a unified platform for Quality Control that attempts to incorporate different QC pipelines and their associated metrics. The QC tools to come out of our work are designed to be easily integrated into MRIQC (Esteban et al., 2016, 2017), a project affiliated to PCPQAP.

Due to the uncertainty about the suitability of the QC metrics discussed above to successfully assess image quality objectively and to detect the majority of brain artefacts, we propose a combination of techniques, including machine learning, to help identify problematic images based on a wide range of imaging derived metrics.

For this study, we combined the generated IDPs with a collection of additional QC-specific metrics (e.g., measures of asymmetry, normalised intensity per subcortical structure, metrics to detect alignment that classify images as usable or non-usable) to detect quality issues. The combination of all these methods is described in Section 3.3. The QC-specific metrics were used in a machine learning tool called autoQC. This has been developed for, and tested on T1 images, and will be further developed in future work so that it can be applied to all modalities. Section 3.3.1 further describes the QC-specific metrics and the machine learning system. The main reason for our heavy focus to date on QC for the T1 data is that the entire processing pipeline (for all modalities) relies heavily on having a

²<http://preprocessed-connectomes-project.org/quality-assessment-protocol/> (Craddock et al., 2016)

usable T1 structural image.

Section 3.3.2 shows different ways to approach the QC process by detecting outliers, while section 3.3.3 describes a structural motion metric used as both a QC feature and a confound. In Section 3.4, we describe the Quality Control performed on FreeSurfer.

First, however, head motion is one of the main issues related to QC and in Section 3.2 we show an investigation we made on the amount of head motion that makes the repetition of a T1 acquisition worthwhile.

3.2 Usability rules

In the first phases of brain image acquisition in UKB, it was noted that a repetition of T1 image acquisition was not infrequent in cases where the radiographer saw that there were some quality artefacts due to head motion. As the acquisition time per subject is fixed, these repetitions imply that the last modalities were not acquired for some subjects.

Not knowing how much these repetitions improved the overall quality and intending to maintain the protocol to be as complete (i.e. acquiring all modalities per subject) as possible, we decided to evaluate whether these repetitions actually improved the quality.

In order to do so, an MRI expert manually evaluated the amount of motion noise in 265 pairs of T1w images (these pairs were repetitions acquired by the scanning teams). The images were manually classified into four categories in terms of motion artefact:

- 1: Good quality.
- 2: Medium quality.
- 3: Bad quality (likely unusable).
- 4: Very bad quality (unusable).

Some examples of categories 2, 3, and 4 can be seen in Figure 3.1. Figure 3.2 shows how repeating the acquisitions improved or worsened the quality for each category. From these results:

- Category 1 (good quality first scan): 7% of the images (good quality) become unusable in the repeat scan. 29% become worse (although not unusable), and 64% stay the same.
- Category 2 (medium / borderline quality first scan): 27% become better in the

repeat scan, 48% stay the same, and 25% of the images become unusable.

- Category 3 (bad quality first scan): 52% of the images become usable in the repeat scan, while 48% remain unusable.
- Category 4 (very bad quality first scan): 60% of the images become usable after repeating the acquisition.

This study helped us reach the recommendation that repeating the acquisition was only advisable in bad motion cases (categories 3 and 4), as repeating the acquisition for images with weak / small artefacts improved the quality of the images with about the same likelihood as getting even worse. These recommendations were passed to the scanning teams and are now in use as part of the standard acquisition protocol.

3.3 QC framework

We developed a QC framework for UK Biobank, to determine whether a modality is unusable (due to bad quality, wrong dimensions, or merely the absence of the modality). The term “unusable” in this context does not imply that the data (if present) will be fully deleted (in part because we wanted this data, that we consider unusable, to be available to researchers, should they wish), but instead, the data inside the modality directory will be moved into a sub-directory called “unusable”, and a text file will be generated including the reasons for this moving. It is also important to mention that making a modality unusable will recursively make unusable any other that depends on it. For example, if the B0 fieldmap³ is deemed unusable, dMRI, rfMRI and tfMRI will be deemed unusable too, as the processing of those modalities depends on the fieldmap. Similarly, making the T1 unusable will make all other modalities unusable, as processing for all of them depends on the T1.

This hierarchy can be seen in Table 3.1:

³Both fieldmap and FreeSufer are considered “modalities” in this context, although strictly speaking, they are not.

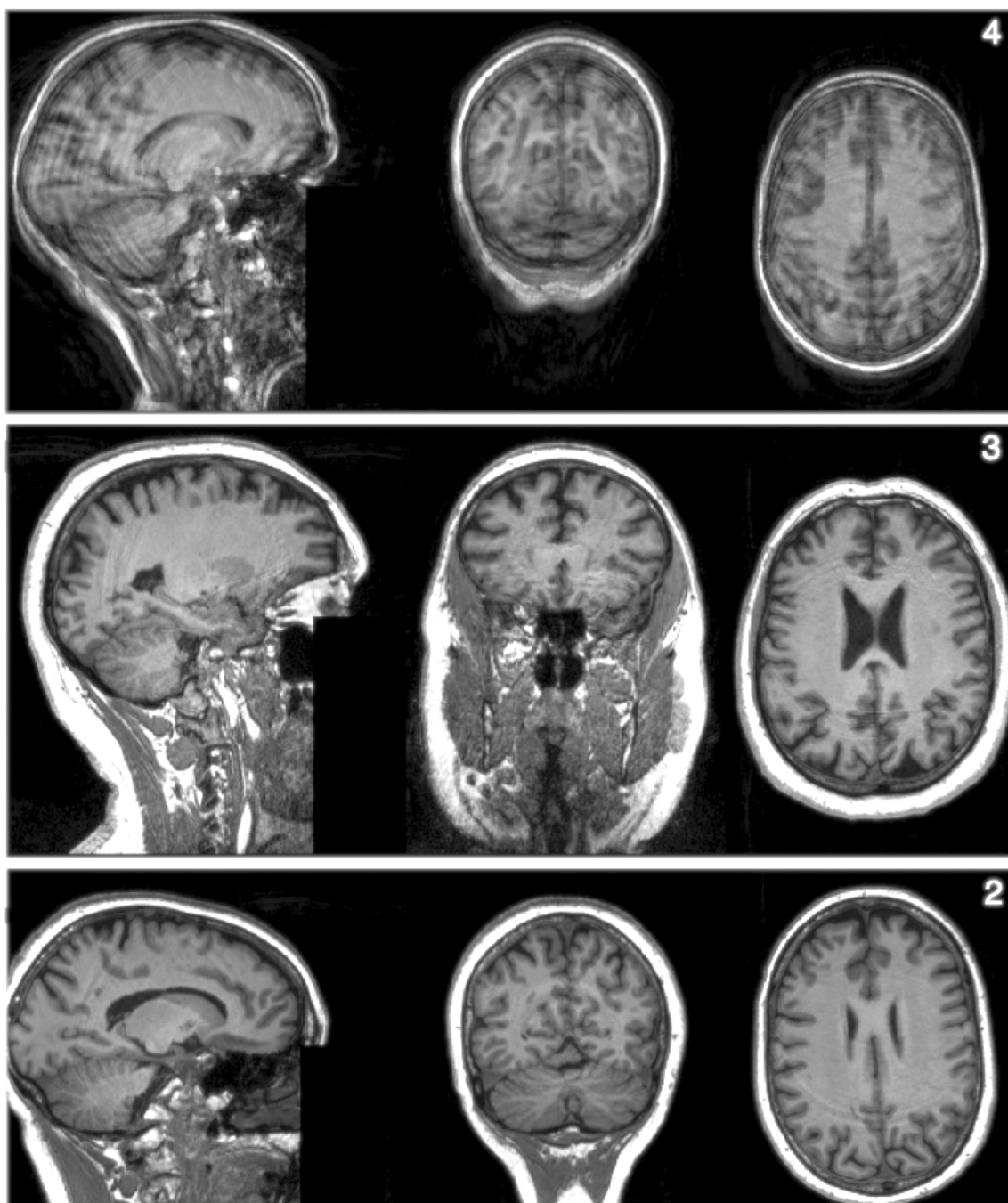


Figure 3.1: Examples of head motion severity for T1w images.
Categories: 4 (very bad quality - unusable), 3 (bad quality - likely unusable) and 2 (medium quality - borderline usable)

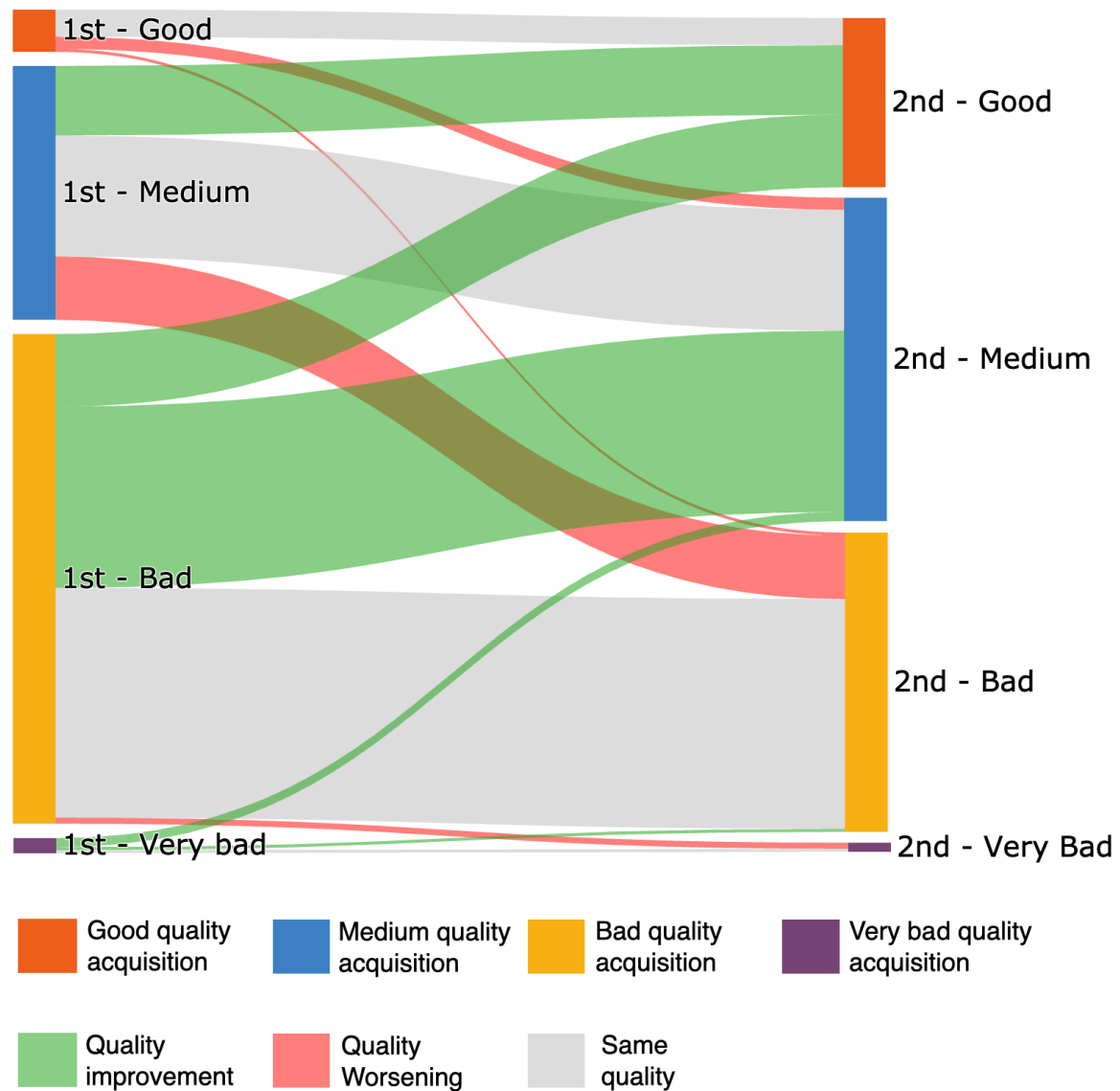


Figure 3.2: Flow plot showing that there is only a useful improvement in quality when the detected artefacts in the first acquisition (left) are really severe, while repeating scan with small artefacts does not on average improve the quality.

Modality	Dependent modalities
T1	Fieldmap, dMRI, rfMRI, tfMRI, T2 FLAIR, SWI, FreeSurfer
T2 FLAIR	None
SWI	None
FreeSurfer	None
Fieldmap	dMRI, rfMRI, tfMRI
dMRI	Fieldmap, rfMRI, tfMRI
rfMRI	None
tfMRI	None

Table 3.1: Hierarchy of modalities for unusability

The rules that would make a modality unusable would be:

- Modality not present (DICOM unavailable).
- Wrong dimensions / resolution / number of images of the modality.
- Overlap between defacing mask and brain mask is greater than 500 voxels.
- Non-background tissue in any of the five border slices (most inferior slices are not checked)⁴.
- Manually verified as “unusable” after being visually checked (if it was in the list of candidates for being unusable - see below).

The general way these rules would apply would be the following: After the main file naming and organisation mentioned in Section 2.2.1.2, a basic Quality Control tool⁵ is used to check if every raw dataset in each modality has the correct dimensions, defined in a configuration file. When raw data has the wrong dimensions, is corrupted, missing, or otherwise unusable, it is moved into a subfolder called “unusable” (inside the given modality’s folder), and not processed any further (apart from defacing applied to the raw T1 and T2 FLAIR).

After this, all the data of all subjects is processed. Following this, the overlap between the defacing mask and the brain mask is checked, and all modalities will be deemed unusable if the overlap is too big.

The next step would be to elaborate a list of “candidates to make unusable”. This list has datasets that meet at least one of these criteria:

⁴This criterion was only applied to T1 images, and due to their small number, they are all manually checked.

⁵[Code for the basic QC tool.](#)

- Flagged by the autoQC (see Section 3.3.1) with normal classes.
- Flagged by the autoQC with binary classes⁶.
- Outliers detected in either QC features or IDPs (see Section 3.3.2).
- Lower than 0.75 Structural Head Motion value (see Section 3.3.3).
- Lower than 55 QoalaT value (see Section 3.4).

After manually checking all subjects in that list of candidates, we separate “actual” unusable datasets from false positives.

These “unusable” datasets are included in the NIFTI packages in the UK Biobank database because some researchers may be interested in working with them, for example, to develop methods for detecting or even possibly correcting such data. However, they are distributed without any further pipeline, processing applied to the unusable raw images, apart from the defacing.

3.3.1 AutoQC for T1

We have developed an automated Quality Control tool to identify T1 images with problems either in their acquisition or in later processing steps. This uses machine learning methods with supervised learning. We first categorise the different problems that we may find in the images and manually classify many datasets accordingly. We then generate a set of QC features aiming to characterise those images. Finally, we feed that information into a supervised learning classifier. With this approach, we aim to have a classifier capable of detecting problematic images with acceptable accuracy. To date, we have focused just on automated QC for the T1 images, as the entire processing pipeline depends on having a usable T1. All data, numbers, and figures in this whole sub-section correspond to our paper (Alfaro-Almagro et al., 2018).

3.3.1.1 AutoQC categories

Table 3.2 shows the categories for QC that were developed to classify issues that may be found in T1 images. This list was compiled as new issues were identified (through manual inspection) in the data during the processing of the first two public releases (batches 1 and 2) and is not intended to cover every possible problem in any MR image, but rather to cover all of the problems that we have observed to date in the Biobank data

⁶AutoQC is run with the same features and the same classifier but two different sets of categories to learn: the normal classes described in 3.3.1.1, and a binarisation of those classes into 0 = good, 1 = bad.

Table 3.2: Classification of problems / imperfections for T1

Problem code	Imperfection code
0 = No problem 1 = Multiple / Unknown problems 2 = Missing or incomplete modality 3 = Bad FOV 4 = Bad registration: Bad head motion / Noise 5 = Bad registration: Structurally atypical 6 = Bad registration: Large Ventricles 7 = Bad registration: General Registration failure 8 = Bad registration: Bad brain mask on the top 9 = Bad registration: Bad brain mask on the temporal lobe 10 = Bad registration: Brain mask out of brain	1 = Multiple / Unknown probs. 2 = Bad head movement 3 = Movement-related blurring 4 = Bias field / contrast problem 5 = Structurally atypical 6 = Problem on top (brain mask) 7 = Problem on temporal lobe (brain mask)

Incomplete = The expected number of images or any of the dimensions (including the temporal dimension) is incorrect.

Structurally atypical: The subject is anatomically atypical. This may be due to a pathology although there was no clinical assessment to confirm this. The problem may be so severe that registration is strongly affected (and thus, the image is deemed unusable for the automated processing pipeline).

set. For this study, we define a “problem” as any issue that makes subsequent processing steps impossible or unreliable, while an “imperfection” is an issue which must be noted, but does not necessarily impede further processing. Some issues (e.g., bad head motion) appear in both categories (“problem” and the less serious “imperfection”), where the distinction is to be made based on the seriousness of the image artefacts. Table 3.3 shows a similar table for all modalities. Figure 3.3 difference in the IDPs between “problematic”, “imperfect” and “good” datasets). Figure 3.4 shows some examples of some datasets from these categories.

New categories were discovered and added for later data releases after (Alfaro-Almagro et al., 2018), including 13: *Deface overlapping*, 20: *Bad dimensions in T1*, and 50: *Subject opted-out of the study*. The number of subjects in each category up to the date of this thesis submission can be seen in Tables 3.12 and 3.13.

Table 3.3: Classification of problems / imperfections per modality in UK Biobank

Modality	Problem code	Imperfection code
T1	4 = Bad reg: Bad head motion / Noise	1 = Multiple / Unknown imperfections 2 = Bad head movement 3 = Movement-related ringing / blurring 4 = Bias field / contrast problem 5 = Structurally atypical 6 = Problem on top (brain mask) 7 = Problem on temp. lobe (brain mask)
	5 = Bad reg: Structurally atypical	
	6 = Bad reg: Large Ventricles	
	7 = Bad reg: General Registration failure	
	8 = Bad reg: Bad brain mask on the top	
	9 = Bad reg: Bad brain mask on the temporal lobe	
	10 = Bad reg: Brain mask out of the brain	
	4 = T1 unusable	
	5 = Bad registration to T1	
	6 = Incompatible (Phase 2)	
T2 FLAIR	4 = T1 unusable	1 = Multiple / Unknown imperfections
	5 = Bad registration to T1	
	6 = Distortion	
	7 = Acquisition problem	
swMRI	8 = Rotated sform / qform	1 = Multiple / Unknown imperfections
	4 = T1 unusable	
	5 = Bad EPI distortions	
	6 = Bad bias field	
dMRI	7 = Bad registration to T1 or MNI	1 = Multiple / Unknown imperfections
	8 = Incompatible (Phase 2)	
	4 = T1 or fieldmap unusable	
	5 = Bad EPI distortions	
rfMRI	6 = Bad bias field	1 = Multiple / Unknown imperfections 2 = Missing/corrupt SBRref
	7 = Bad registration to T1	
	8 = Bad acquisition (Movement)	
	(Same as for rfMRI)	
tfMRI	(Same as for rfMRI)	(Same as for rfMRI)

Problem codes for all modalities:

- 0 = No problem
- 1 = Multiple / Unknown problems
- 2 = Missing or incomplete modality
- 3 = Bad FoV

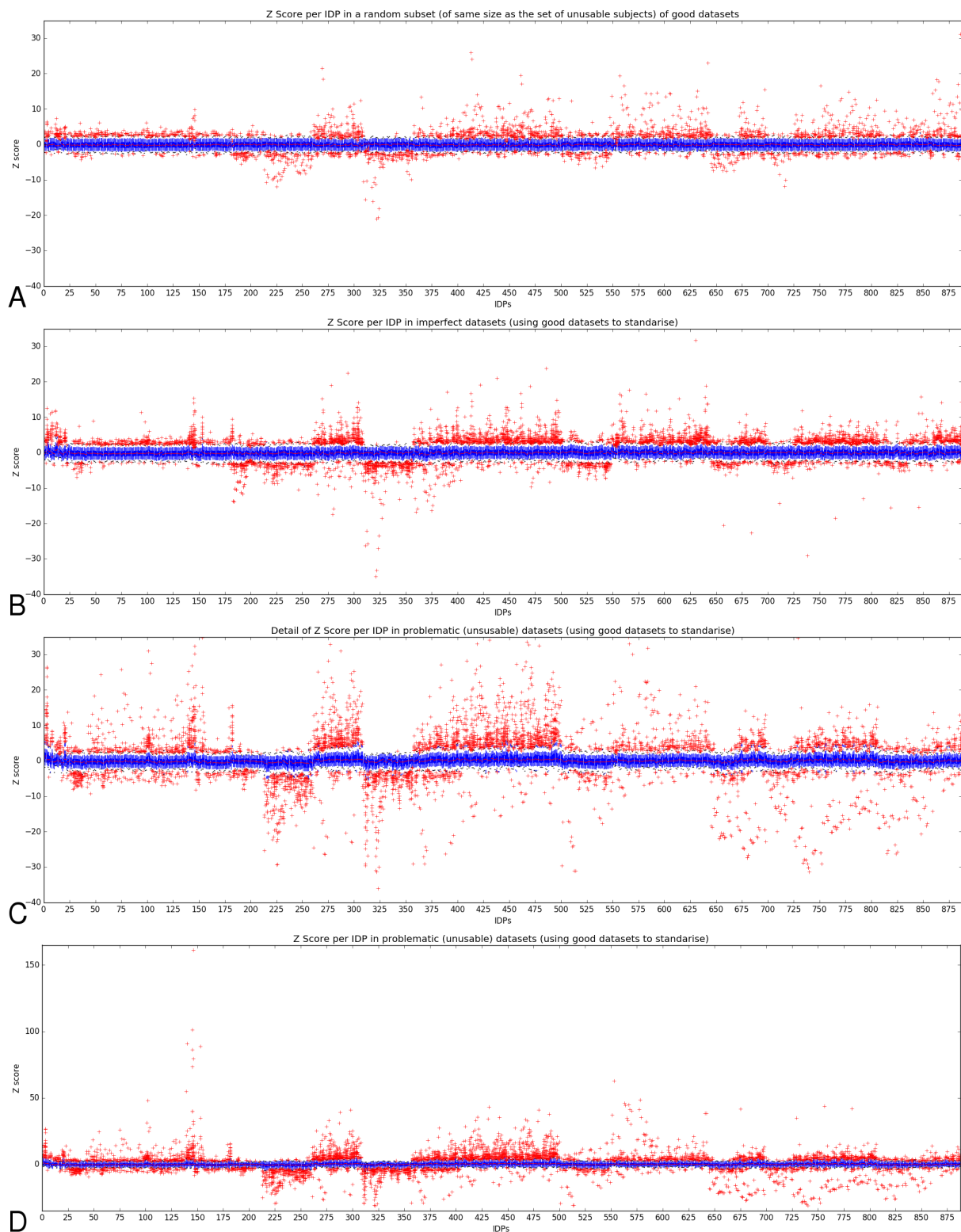


Figure 3.3: Boxplots with the Z scores for the 888 not-rfMRI IDPs (i.e. excluding netmats, and listed in the same order as described above) for *A*: Good datasets, *B*: Imperfect datasets, *C* and *D*: Unusable (“problem”) datasets. Y axis has the same range for A, B and C to ease comparison. Full range for unusable datasets is shown in D. Blue boxes show the IQRs per IDP. Red points above and below are the boxplot-defined outliers.

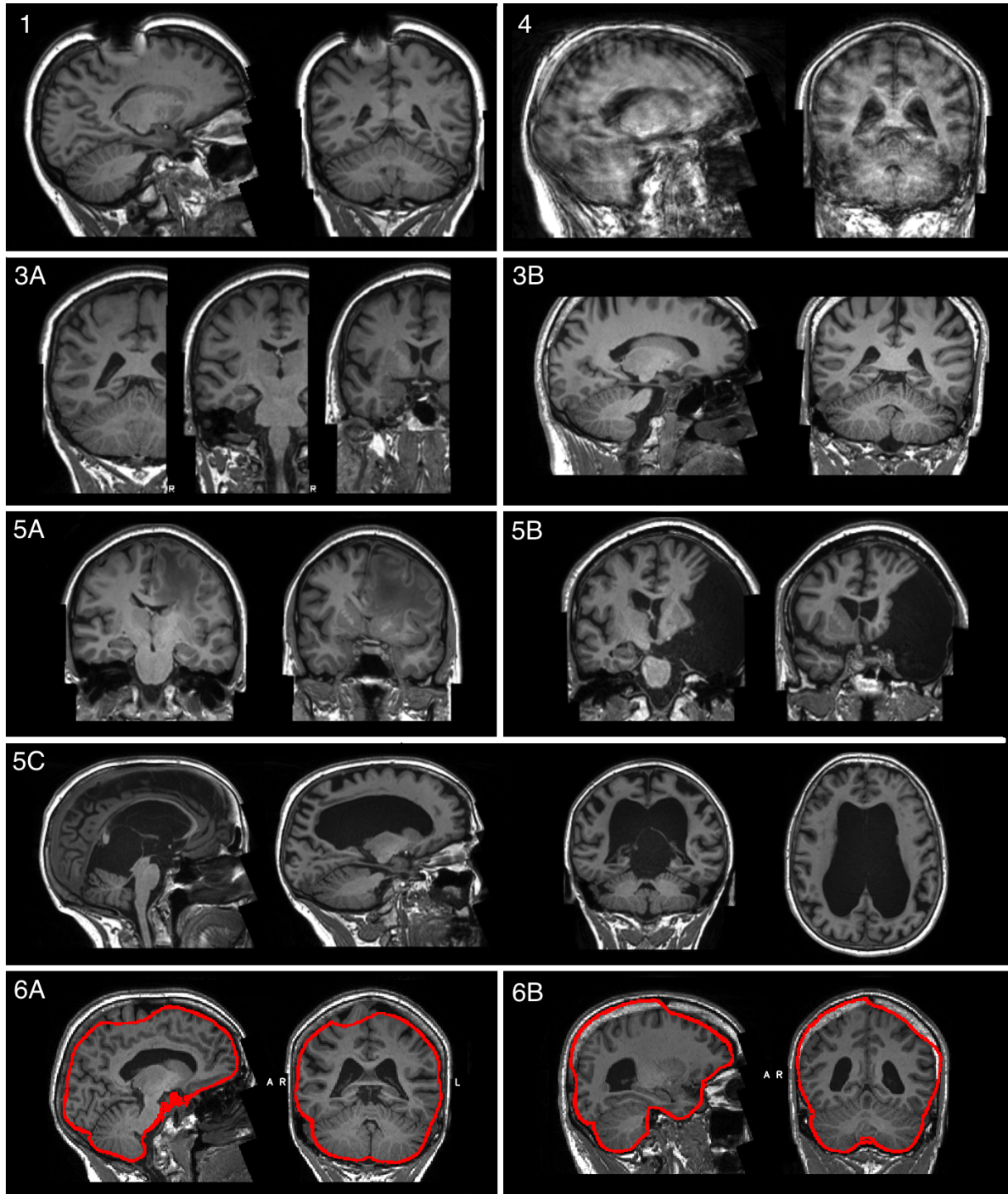


Figure 3.4: Examples of problematic (“unusable”) datasets. Numbers correspond to the classes described in Table 3.3. *Class 1*: Multiple or Unknown problems. *Class 4*: Bad head motion. *Class 3 (A and B)*: Bad FoV. *Class 5 (A, B and C)*: Structurally atypical. *Class 6 (A and B)*: Bad registration: Unknown reason

3.3.1.2 AutoQC features

We developed a set of 190 features (for use in the trained machine learning automated classifier) to find issues in T1 structural images. Some of these QC features (the first 31)

are also IDPs for T1 images as listed 2.1.1 above. We have divided these features into 17 groups:

- The first three features (QC-T1-001 to QC-T1-003) describe the discrepancy between the T1 structural image and the standard (population average) template image. This discrepancy (generalised difference) implies that an extreme score could arise either because the quality of the alignment is bad, or because one of the two images being compared is different in some way, either due to artefacts or biological differences. The three metrics are the discrepancy after linear alignment (which cannot correct for fine-grained differences between a given subject and the population average); the discrepancy after non-linear alignment (which should achieve much better correspondence); and the overall amount of non-linear warping (utilising the Jacobian of the non-linear warp field) necessary to achieve a detailed alignment to the standard template.
- The next two features (QC-T1-004 and QC-T1-005) measure signal to noise ratio (SNR) of the T1. Brain and tissue-type segmentation⁷ is used to estimate background noise level (standard deviation), as well as mean intensities for grey and white matter. These quantities are used to estimate overall image SNR and also CNR (contrast to noise - white-grey mean intensity difference normalised by noise level). In both cases, these measures are inverted before being recorded as QC measures, so that higher is worse. As with the previous features, similar measures can be calculated from preprocessed rfMRI (both before and after artefact removal) and tfMRI time-series data. In this case, however, the “noise” level is the temporal standard deviation.
- The next eleven features (QC-T1-006 to QC-T1-016) show SIENAX outputs (see Section 2.2.2) with the normalised and non-normalised volumes for the whole brain, and for the segmented tissue types. These metrics can be useful in identifying problems in the T1 acquisition as well as the registration and brain extraction, as the segmentation strongly depends on the brain extraction.
- The next fifteen features (QC-T1-017 to QC-T1-031) show the volumes of the 15 subcortical structures identified in the T1⁸. Extreme cases of these features can come either from acquisition or processing problems, or from variability that may

⁷Output of the FAST tool (Zhang et al., 2001)

⁸Output of the FIRST tool (Patenaude et al., 2011)

be worth studying (and therefore may not represent a problem in the processing, but rather an anatomical variability).

- The next seven features (QC_T1_032 to QC_T1_038) show the asymmetry defined as the ratio of larger volume / smaller volume between the seven paired subcortical structures from the previous block.
- The next eight features (QC_T1_039 to QC_T1_046) show the global asymmetry in the brain calculated as the difference between the brain-extracted T1 and the same image flipped in the left-right axis (after a prior linear registration to the MNI152 template and normalisation with the median intensity of the T1). The eight metrics cover various descriptive statistics calculated from this difference (such as maximum value, mean value, etc.). A very strong global asymmetry can be due to extreme structural variability (which is one of the cases described in Section 3.3.1.1) or problems in acquisition or processing.
- The next sixty features (QC_T1_047 to QC_T1_106) contain the mean and standard deviation of the intensity of the 15 main subcortical structures normalised using two different methods:
 - Method 1: Subtracting the mean intensity in the grey matter from the mean intensity in the structure and dividing by the difference between mean intensity in white matter and mean intensity in grey matter.
 - Method 2: Dividing mean intensity in the structure by mean intensity in grey matter.
- The next feature (QC_T1_107) contains the volume of grey matter that lies outside of the subject's brain mask. This grey matter volume was calculated with FAST on a brain extracted T1 using BET (As opposed to the brain extraction method described in Section 2.2.2). The idea behind this feature would be to be able to detect grey matter out of the brain mask calculated by our pipeline using an alternative brain extraction method.
- The next nine features (QC_T1_108 to QC_T1_116) contain the amount of all tissues from the FAST segmentation that lie on the edges of the subject's brain mask. An example of this can be seen in Figure 3.5. These features could be useful for identifying problems in brain extraction and registration, as the amount of white matter should be negligible on the edges of the brain (the edges should primarily consist of grey matter / CSF).

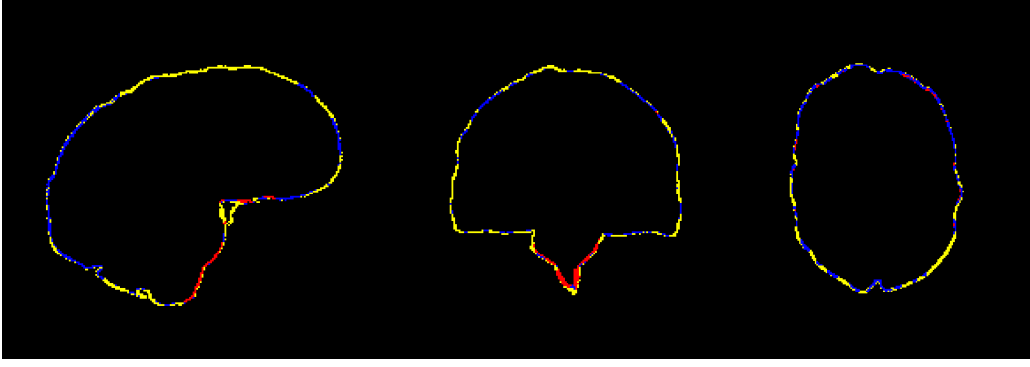


Figure 3.5: Volume of different tissues on the edges of the standard brain mask. *Red*: White matter ($2,389 \text{ mm}^3$). *Blue*: Grey matter ($44,696 \text{ mm}^3$). *Yellow*: CSF ($62,335 \text{ mm}^3$).

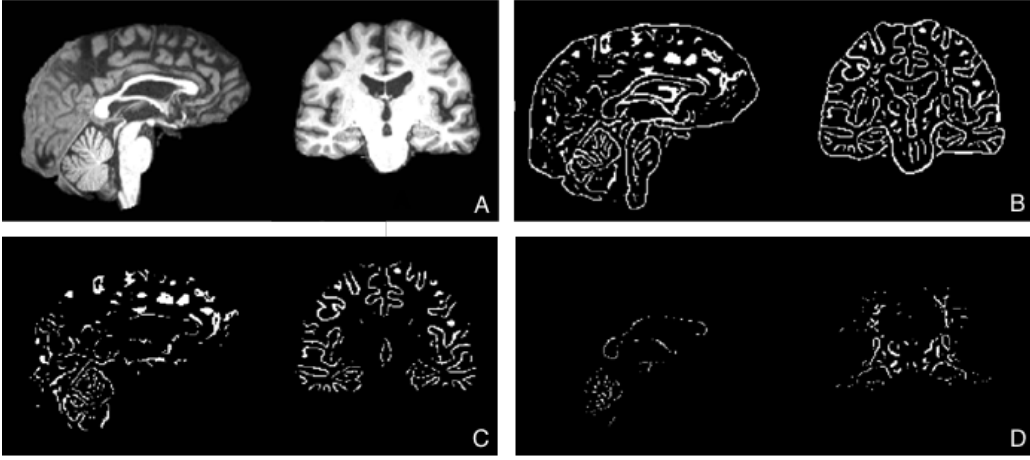


Figure 3.6: Volume of edges identified using a Canny filter on different tissue types. *A*: Brain extracted T1. *B*: Edges in the whole brain. *C*: Edges in the GM using FAST segmentation and cleaning edges on the brain mask boundary, ventricles and brain stem. *D*: Edges in the WM using the same cleaning as in C.

- The next seven features (QC-T1_117 to QC-T1_123) contain the volume of edges calculated with a Canny filter on the brain extracted T1. Edges that lie on the brain mask boundary, the ventricles and the brainstem are excluded. An example is shown in Figure 3.6. The utility of these features is very similar to the tissue edge features described above.
- The next eleven features (QC-T1_124 to QC-T1_134) describe a comparison (using the Dice coefficient in a similar manner to [Kazemi and Noorizadeh \(2014\)](#)) between the Biobank brain mask (as described in Section 2.2.2) and other brain extraction methods, including:
 - Non linear registration to a study specific template (As opposed to the MNI152 template).
 - *mriwatershed* method from Freesurfer ([Ségonne et al., 2004](#)) using default pa-

rameters.

- mriwatershed method from Freesurfer using the atlas option.
- ROBEX (Iglesias et al., 2011).
- SPM (Ashburner and Friston, 2005).
- BSE from BrainSuite (Shattuck and Leahy, 2002) with default parameters.
- BSE from BrainSuite with optimised parameters⁹.
- BET (Smith, 2002) on T1 image.
- BET (Smith, 2002) on T2 FLAIR image.
- Linear registration of the MNI brain mask to T1 space.
- Mask created by selecting the voxels that belong to a majority (Number of available masks -2) of the previously created masks.

With these features, we are attempting to avoid the circularity of evaluating the quality of Biobank brain extraction by using features which depend upon that brain extraction.

- The next six features (QC_T1_135 to QC_T1_140) contain six descriptive statistics (mean, median, and 2nd, 25th, 75th, and 98th percentile) of the 11 previously mentioned DICE coefficients for the brain extraction methods described above.
- The next twelve features (QC_T1_141 to QC_T1_152) contain three descriptive statistics (mean, median and 95th percentile) of the intensity of the internal and external boundary of the brain mask. We calculated the internal and external boundaries on the whole brain mask and the upper half of the brain mask (As that is the portion of the brain with the larger number of problems in the registration). See Figure 3.7.
- The next two features (QC_T1_153 and QC_T1_154) contain the volume of the largest hole in the grey matter and white matter (With a “hole” defined as a cluster of voxels with intensity below the 5th percentile). The largest hole is calculated using a Z threshold of 0.5.
- The next five features (QC_T1_155 to QC_T1_159) contain some descriptive statistics (Mean, median, 75th and 99th percentile, and the number of voxels with extreme values¹⁰) of the magnitude of the non-linear registration to the MNI152 template

⁹Diffusion constant = 50; Diffusion iterations = 1 ; Edge detection constant = 0.7 (-d 50 -n 1 -s 0.7). These parameters were optimised with a grid search, using similarity with Biobank brain mask as the cost function

¹⁰Mean + 3 times the standard deviation

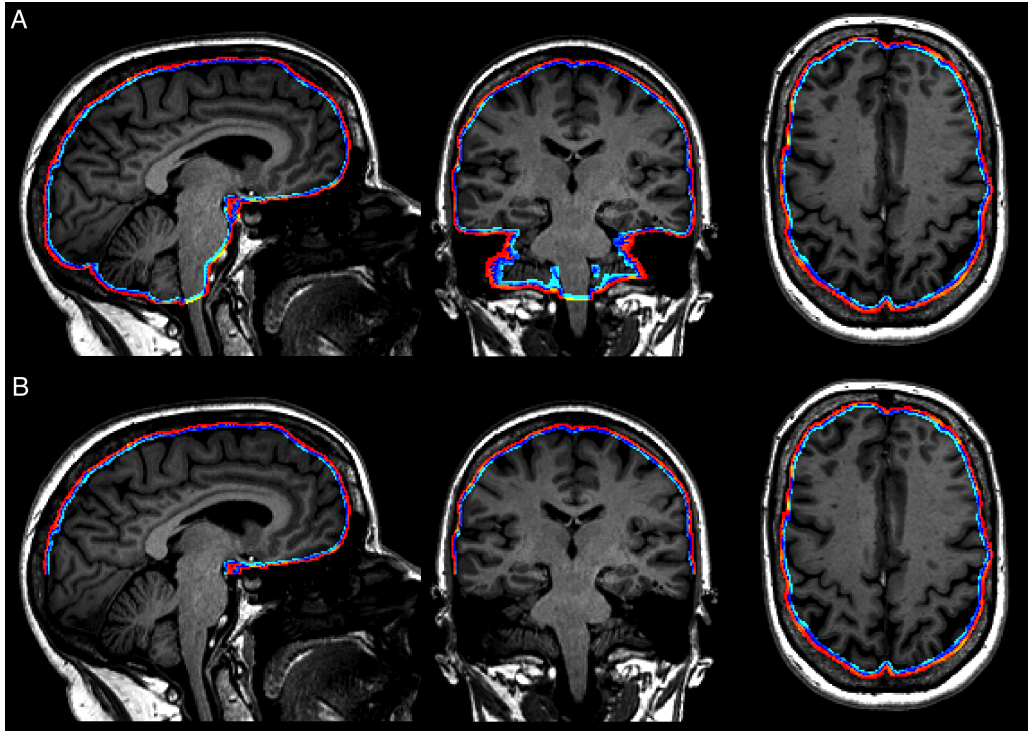


Figure 3.7: Interior (Blue) and exterior (Red) strip boundary of the brain mask. *A*: Whole brain mask. *B*: Upper half brain mask.

calculated as the square root of the sum of the three squared dimensions of the estimated warp field.

- The next feature (QC_T1_160) contains the volume of white matter hyper-intensities in structural images using T1 and T2 FLAIR as inputs to BIANCA ([Griffanti et al., 2016](#)). This feature may be used to detect atypical structures that could affect registration.
- The next thirty features (QC_T1_161 to QC_T1_190) contain descriptive statistics (minimum, maximum, median, and 75th and 98th percentile) of the distance between the boundaries of 6 different structures (cerebellum, parietal lobe, temporal lobe, occipital lobe, frontal lobe and total brain), and the boundary of the MNI152 brain mask. Figure 3.8 shows an example with temporal lobe. These features may be useful in identifying atypical deformations due to poor registration to MNI152 space.

3.3.1.3 AutoQC tool description

We used the Weka machine learning toolbox ([Hall et al., 2009](#)), with three separate classifiers' outputs fused together. The ensemble classifier used for the fusion was a voting system that combines the a posteriori probabilities of the different classifiers using

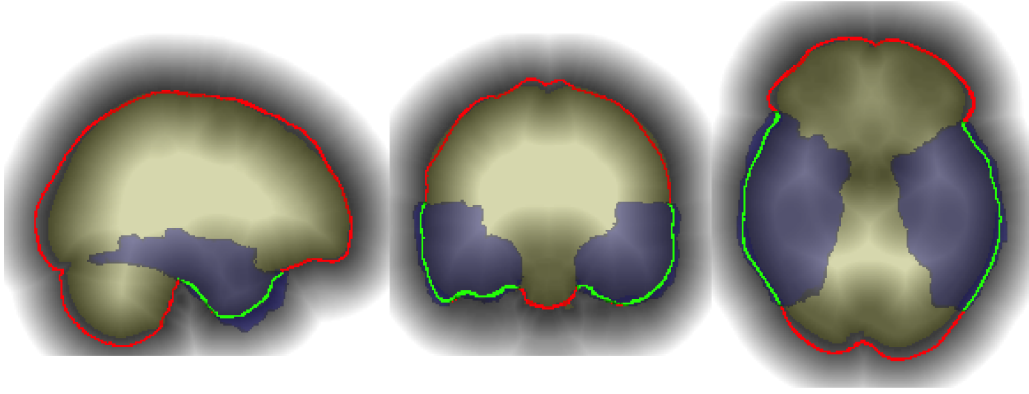


Figure 3.8: Calculation of the distance between the border of the temporal lobe and the border of the MNI template brain mask. *Greyscale*: Distance map of a voxel to the border of the MNI template brain mask. Higher (brighter) intensities mean greater distance. *Yellow*: Outline of the MNI template brain mask. The previous distance map has 0s in the voxels in the border of this mask. *Red*: Outline of the brain mask of the subject taken to MNI space with a linear (12 DOF) registration calculated between the T1 and the MNI T1 template. *Blue*: Dilated mask of the temporal lobes of the subject (Calculated by applying the inverse non-linear registration from T1 to MNI described in Section 2.2.2 to a mask of the temporal lobe defined in MNI space) taken to MNI space using the same linear registration mentioned before. *Green*: Intersection between blue and red. Distance maps defined in greyscale will be masked with this ROI to calculate a set of statistics.

the ‘Minimum Probability’ combination rule (Kuncheva, 2004; Kittler et al., 1998). This rule was chosen¹¹ among five different options based on our goal of reducing the FNR (False Negative Rate: rate of missing “problems”) as much as possible without increasing the FPR (False Positive Rate: rate of incorrectly flagging “good” cases as “problems”) to impractical levels¹². The three combined classifiers were:

- Bayes Network classifier (Bouckaert, 2008): This classifier first finds the Bayesian network (as a directed acyclic causal graph) that best fits the data by using the training dataset to create the network structure (running conditional independence tests to find the causal structure of the network, using local score metrics on the quality of nodes/edges, and performing global score metrics by estimating classification accuracy) and then assigning probabilities to each node using the distribution of the same training dataset; the inference is later performed by using the maximum a posteriori decision rule for the node of interest (i.e., the variable being predicted) for a new instance (a vector of features).
- Naive Bayes classifier (John and Langley, 1995): This classifier is a simpler version of

¹¹By testing on a reduced version of the final training set

¹²As all the cases flagged as “problems” would need a posterior visual inspection

	0	1	3	4	5	8	9	10
0	0	15000	18000	15000	15000	15000	18000	15000
1	30000	0	1	1	1	1	1	1
3	5000000	1	0	1	1	1	1	1
5	30000	1	1	0	1	1	1	1
6	30000	1	1	1	0	1	1	1
8	3000000	1	1	1	1	0	1	1
9	2000000	1	1	1	1	1	0	1
10	5000000	1	1	1	1	1	1	0

Table 3.4: Cost matrix represents the hypothetical cost of misclassifying a subject belonging to the category X (columns) in the category Y (rows). Note that the categories correspond to those detected for batches 1 and 2 in Table 3.13.

the above one in the sense that it assumes strong (naive) independence between the features (in the sense that there is no correlation between the different features) and therefore does not need to generate a graph structure. The maximum a posteriori decision rule for the parent node is also used here for inference.

- MetaCost classifier (Domingos, 1999): MetaCost is a meta-classifier that allows the use of different costs for penalising (e.g.) FPR vs FNR. for a certain base classifier. In our case, as the primary goal was to try to minimise the False Negatives (subjects that are actually unusable but were classified as usable) we created a cost matrix that penalised this kind of misclassification greatly. This cost matrix was calculated using a grid search of different parameters and can be seen in Table 3.4. The base classifier was another Bayes Network classifier.

Combining classifiers usually tries to compensate for possible overfitting (or lack of fitting) of individual classifiers. The reason for choosing this combination of algorithms was that they allowed us to weigh the importance of type I vs type II errors, besides exhibiting a satisfactory accuracy in earlier stages of the classification (again, with a reduced version of the current training set).

3.3.1.4 Performance of autoQC

In order to validate our QC system, image quality for 5816 subjects was manually assessed. Problems were identified in 103 subjects. Table 3.5 shows the performance of the classifier in a stratified¹³ 10-fold cross-validation.

¹³The proportion of the classes to classify is maintained in the folds, which is important in unbalanced situations

	Positive	Negative	
Classified Positive	TP: 94	FP: 912	P. Precision: 9%
Classified Negative	FN: 9	TN: 4801	N. Precision: 99.8%
	Sensitivity: 91%	Specificity: 84%	

Positive = Unusable datasets. **Negative** = Usable datasets.

Datasets flagged for manual review: 1006 / 5816 (17.29 %)

Unusable datasets missed: 9 / 5816 (0.15 %)

Table 3.5: Automated QC tool performance for T1 images. Note that these are the numbers reported for (Alfaro-Almagro et al., 2018). The numbers have changed, as can be seen in Section 3.5, after improving the whole QC process as described in Section 3.3 and modifying the QC categories as described in Section 3.3.1.1.

The majority of problems found in the processing are related to non-linear registration, which likely therefore also affect the brain extraction. The 103 subjects were deemed unusable due to irreconcilable problems with non-linear registration, in many cases caused by poor data quality (for example, where the T1 is corrupted by bad subject head motion, as can be seen in panel 4 of figure 3.4). The QC system we have developed aims to detect problems like this.

The number of false negatives (that is, subjects that are actually unusable but were classified as usable), is low (9 subjects, i.e. 0.15%), though in future work we will aim to reduce this even further. Choosing parameters/thresholds to drive this number of false negatives so low comes at the cost of a much higher false-positive rate ($912 / 5713 = 16\%$). Nevertheless, these results guarantee a considerable reduction in the number of subjects in future data sets that will then require manual QC assessment (i.e., manually checking 1006 subjects instead of 5816).

This automated QC tool was used on the second release of 4282 subjects, significantly reducing the manual checking to approximately 750 subjects, resulting in the detection of 71 unusable datasets.

3.3.1.5 Size of the training dataset for QC

The classifier accuracy in detecting problematic datasets will naturally improve with increasing training set size (as the classifier will have more information to model the patterns that describe the classes). In order to obtain a somewhat quantitative proof of this idea,

we have re-run the classifier accuracy estimation by separating a small subset of the original 5822 subjects (1000) and training the classifier with a progressively smaller fraction of the other 4822 datasets (4770, 4721, etc, until just having 48 in the last iteration).

As shown in Figure 3.9, a reduction in the size of the training set marginally improves the number of incorrectly classified datasets in the test set (and also reduces the amount of datasets to manually review). Nevertheless, this small reduction comes at a cost: An increase in the problematic datasets that are missed by the classifier. The explanation for this behaviour is that the less bad subjects to learn from in the training set, the less likely it is to classify a bad subject correctly. Therefore, it is more likely that most of the subjects will be classified as “good”. As the proportion of bad:good is very small, even classifying all subjects as good (which is what the classifier is doing in practice with the smallest training set sizes tested here) will have very good accuracy, although the QC tool is then of course not functioning usefully.

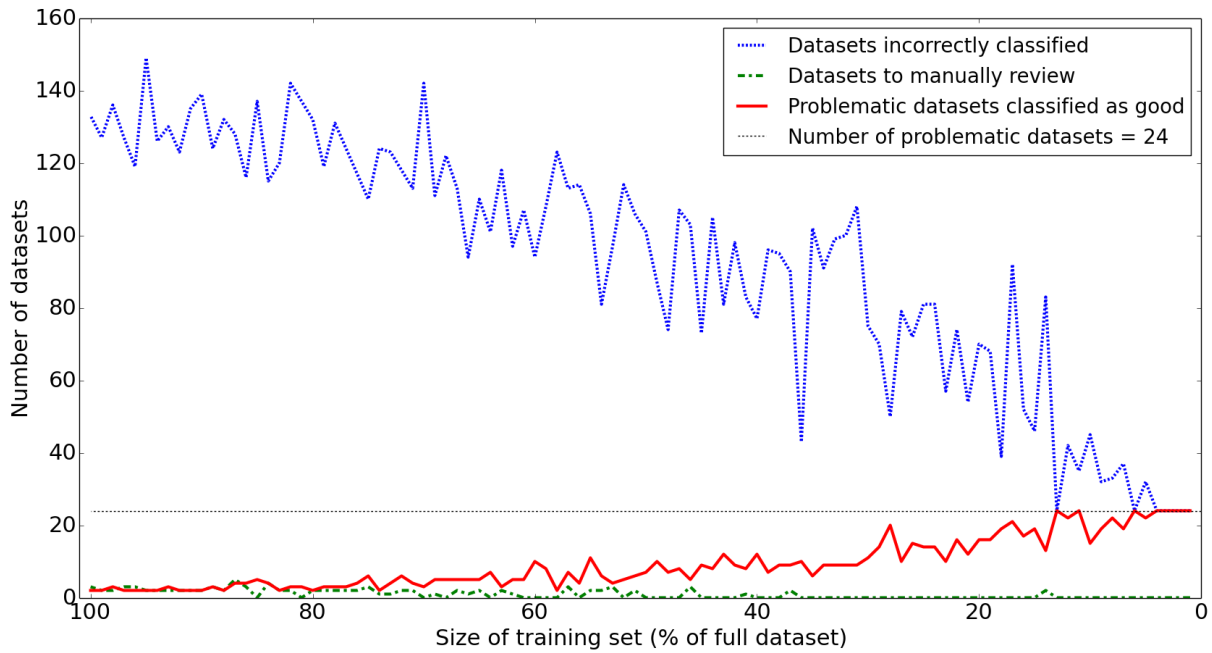


Figure 3.9: Increase in False Negatives when training size decreases.

3.3.2 Outlier detection for QC

A dataset can enter into the list of candidates for unusability if either the T1 IDPs or the QC features contain outliers. This outlier detection is based on the calculation of the z-score¹⁴ and setting a threshold to determine what constitutes an outlier.

¹⁴Subtracting the mean and dividing by the between-subject standard deviation, with both statistics having been calculated from usable datasets.

To this end, we decided that any value beyond ± 3 would be an outlier, and any value beyond ± 5 would be an extreme outlier.

A dataset would enter the list of candidates (to be manually checked) if it either contains an extreme outlier or if it contains a number of outliers:

- The dataset has at least one extreme outlier ($\text{abs}(\text{Z-score}) > 5$) QC feature.
- The dataset has at least 10 outliers ($\text{abs}(\text{Z-score}) > 3$) QC features.
- The dataset has at least one extreme outlier ($\text{abs}(\text{Z-score}) > 5$) T1 IDP.
- The dataset has at least 5 outliers ($\text{abs}(\text{Z-score}) > 3$) T1 IDPs.

3.3.3 Structural motion metric

In order to obtain an estimation of structural motion (i.e. motion from T1w images), we performed a (cross-validated) linear regression where the dependent variable was a manually evaluated QC measure of motion, and the independent variables were a set of features that are related to structural motion.

3.3.3.1 Dependent variable: Manual measure of motion

The measure of head motion was the manual evaluation done by an MRI expert on the 530 T1w images (265 pairs) mentioned in Section 3.2. Figure 3.1 above shows examples of these categories. 355 images manually evaluated as good quality (no motion artefact) were added to balance the classes to a total of 871 images. These quality categories (1: Good quality, 2: medium quality, 3: bad quality, and 4: very bad quality) were converted to a $[0, 1]$ numeric range as described in table 3.6.

Description	Category	Value
Good quality	1	0.00
Medium quality	2	0.75
Bad quality	3	0.90
Very bad quality	4	1.00

Table 3.6: Numeric value for QC category

3.3.3.2 Independent variables

The variables that were calculated to fit the linear regression are associated with noise:

- Estimation of smoothness in dimensions X, Y, and Z using FSL' smoothest ([Flitney and Jenkinson, 2000](#))

- Cost function of the linear registration of the T2-Flair to the T1w
- CNR¹⁵ of the scanner-normalised T1w with the noise defined in the four top corners of the image.
- CNR of the scanner-normalised T1w with the noise defined in the four bottom corners of the image.
- CNR of the not-normalised T1w with the noise defined in the four top corners of the image.
- CNR of the not-normalised T1w with the noise defined in the four bottom corners of the image.
- CNR of the not-normalised T1w with the noise defined in two cuboids defined right in front of the eyes.
- Qoala-T score of the FreeSurfer output (Qoala-T was trained with UK Biobank data),
- Average (across hemispheres) of the Euler number of FreeSurfer surfaces before correction.

3.3.3.3 Linear regression - Fitting and validation

To fit the linear regression, we had 871 T1w images manually labeled by an MRI expert. The number of subject for each category is:

Category	Number of subjects
1	407
2	199
3	253
4	12

Table 3.7: Number of subjects per QC category

To perform the regression, we left-out 20% (174 subjects) of the dataset and tuned the MSE for the remaining 80% (697 subjects) in the following way:

We fitted 21 different regression models from scikit-learn (Pedregosa et al., 2011) (default parameters) with a 10-fold stratified cross-validation and picked the algorithm with the best performance (minimum MSE). These are the results of that process:

¹⁵Contrast-to-Noise-Ratio defined as the contrast between grey matter and white matter compared to noise (within background) standard deviation

MSE for base.LinearRegression:	0.1164
MSE for bayes.ARDRegression:	0.1169
MSE for bayes.BayesianRidge:	0.1165
MSE for coordinate_descent.ElasticNetCV:	0.1167
MSE for least_angle.LarsCV:	0.1168
MSE for coordinate_descent.Lasso:	0.1782
MSE for coordinate_descent.LassoCV:	0.1168
MSE for least_angle.LassoLars:	0.1782
MSE for least_angle.LassoLarsCV:	0.1168
MSE for least_angle.LassoLarsIC:	0.1165
MSE for coordinate_descent.ElasticNet:	0.1782
MSE for omp.OrthogonalMatchingPursuit:	0.1347
MSE for omp.OrthogonalMatchingPursuitCV:	0.1193
MSE for passive_aggressive.PassiveAggressiveRegressor:	0.3796
MSE for ridge.Ridge:	0.1164
MSE for ridge.RidgeCV:	0.1164
MSE for stochastic_gradient.SGDRegressor:	0.1166
MSE for theil_sen.TheilSenRegressor:	0.2163
MSE for ransac.RANSACRegressor:	8.0650

Table 3.8: MSE for the tested regression methods

The algorithm that showed best performance was the basic Linear Regression, with an MSE of 0.1164. The learned weights with this process were:

smoothestX	0.004
smoothestY	0.014
smoothestZ	-0.045
T2_FLAIR_align_to_T1	0.040
T1_orig_QC_CNR_upper	0.055
T1_orig_QC_CNR_lower	-0.029
T1_notNorm_QC_CNR_upper	-0.071
T1_notNorm_QC_CNR_lower	0.027
T1_QC_CNR_eyes	0.092
QoalaT score	0.030
avg_euler_number	-0.09

Table 3.9: Learned weight per feature

When using the learned weights on the hold-out set, the MSE was 0.135. These weights were then applied to the whole dataset to get the metric.

3.3.3.4 Further validation

In order to further validate this metric, we performed a further series of tests. First, we evaluated how the new metric performed on subjects that had originally been deemed unusable by the UK Biobank processing pipeline, using a different quality evaluation method (the semi-automated quality control described in [Alfaro-Almagro, 2018]). As can be seen (Figure 3.10), the proposed metric evaluated as “bad” or “very bad” quality almost all of the T1 images that were previously deemed unusable.

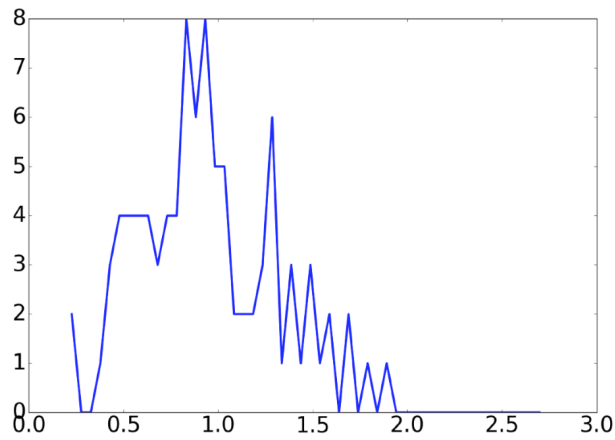


Figure 3.10: Histogram for the structure motion for the subjects deemed unusable due to motion artefact (Alfaro-Almagro et al., 2018). The QC values (encoding level of artefact) in the training data ranged from 0:1 (Table S21), but here the predicted values range above 1, presumably reflecting the extreme corruption (compared with training data) of previously discarded “unusable” data.

Next, we manually checked the 200 more extreme values of the metric (the lowest 100, meaning that the subjects should not have any motion artefact, and the highest 100, meaning that the subject should have significant motion artefact). Two out of the lowest 100 showed a slight motion noise. The rest of them were clean. Ten out of the lowest 100 did not show any noticeable motion noise.

Finally, we checked the correlations between IDPs (plus some QC metrics and some variables later used as a confounds) and the new motion metric to see whether the new metric (intended to be a new confound variable) contained any signal of relevance. Figure 3.11 shows a histogram of those correlations.

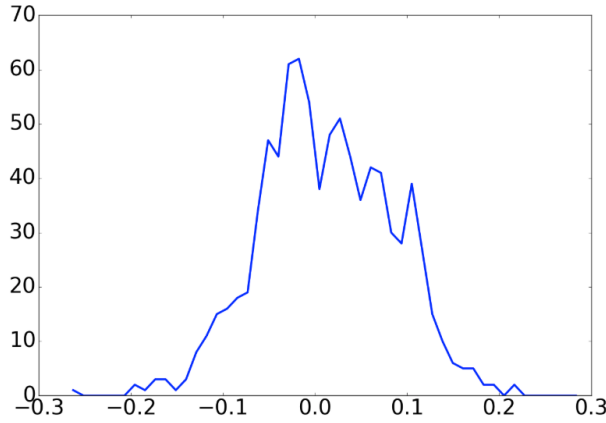


Figure 3.11: Histogram correlations of all non-connectivity IDPs (892 IDPs) with the new metric.

The 40 IDPs / QC metrics / confound variables most strongly correlated (negatively or positively) with the metric are:

BrainBackY	-0.268	BrainCogZ	0.289
T1_SIENAX_headsize_scaling	-0.199	T1_inverse_CNR	0.273
T1_SIENAX_CSF_unnorm_vol	-0.193	T1_nl_align_to_std_warp	0.259
dMRI_TBSS_L1_14	-0.187	tfMRI_align_to_T1	0.255
dMRI_TBSS_L1_12	-0.179	tfMRI_head_motion	0.254
dMRI_TBSS_L1_13	-0.173	T1_GM_parcellation_Cerebellum_vol	0.22
dMRI_TBSS_MO_48	-0.171	dMRI_TBSS_ICVF_48	0.216
dMRI_TBSS_L1_10	-0.162	rfMRI_head_motion	0.206
dMRI_TBSS_MO_10	-0.16	rfMRI_cleaned_TSNR	0.196
tfMRI_p90_zstat_faces-shapes_amygdala	-0.158	dMRI_TBSS_ISOVF_22	0.194
SWI_T2star_l_amygdala	-0.146	dMRI_TBSS_OD_1	0.193
dMRI_TBSS_MD_14	-0.140	dMRI_TBSS_OD_10	0.185
tfMRI_median_zstat_faces	-0.136	dMRI_TBSS_ISOVF_23	0.182
T1_FIRST_left_accumbens	-0.135	dMRI_autoPtx_ISOVF_2	0.175
dMRI_TBSS_FA_39	-0.134	T1_l_align_to_std_diff	0.175
dMRI_TBSS_MO_1	-0.133	TablePosition	0.175
SWI_T2star_l_hippocampus	-0.133	dMRI_TBSS_ISOVF_3	0.174
tfMRI_p90_faces	-0.132	dMRI_TBSS_ICVF_9	0.173
T1_GM_parcellation_R_Amygdala_vol	-0.127	dMRI_autoPtx_ISOVF_3	0.173
dMRI_TBSS_FA_38	-0.125	dMRI_autoPtx_ICVF_18	0.166

Table 3.10: Variables most strongly correlated with the new metric

As can be seen, many of the top IDPs correlated with the new metric are already confounds (BrainBackY, BrainCogZ, T1_SIENAX_headsize_scaling, etc.), or are QC IDPs (T1_inverse_CNR, alignment measures, etc).

We concluded that the metric for structural motion could be useful as a confound variable.

3.4 QC on FreeSurfer

Due to the distinct nature of FreeSurfer (Dale et al., 1999) outputs¹⁶, and their widespread use, we needed to find an alternative way to check the quality of those outputs. In Section 3.4.1, we describe how we have performed this quality control. In Section 3.4.2, we describe the results of the FreeSurfer execution. Finally, in Section 3.4.3, we explain how we run FreeSurfer using either T1 + T2 FLAIR or just T1 for the cases where T2 FLAIR was not available, and the differences between those executions.

3.4.1 Description

Quality Control on FreeSurfer outputs was performed using the tool Qoala-T (Klapwijk et al., 2019), a supervised machine learning tool using random forests as the classification algorithm and some IDPs from FreeSurfer as the learning features.

As with any supervised classification tool, we needed to provide Qoala-T with an already classified training set. An option would have been to use the pre-trained data that comes with Qoala-T, but the idiosyncrasies of the UK Biobank project may not match very well with that training dataset. Therefore, we used the QC labels (and FreeSurfer features) of 4,230 subjects provided by an external group that analysed (and manually evaluated) UK Biobank data independently. We are grateful for these QC contributions from Simon R. Cox, Xueyi Shen, Lianne M. Reus, Clara Alloza, Mathew A. Harris, Helen L. Alderson, Stuart Hunter, and Emma Neilson.

The output of Qoala-T is a value between 0 and 100, showing the probability of the dataset being usable (0% means an absolute certainty that the dataset should be excluded, and 100% means an absolute certainty that the dataset is OK). The official recommendations for the tool would be to manually check any dataset with a probability between 30% and 70%, but the vast number of subjects in UK biobank would make the task unfeasible (we would need to manually check 4,663 datasets out of the 41,621 that were evaluated with QoalaT). Therefore, we reduced the manual-check interval to subjects between 45% and 55%, allowing us only to need to check 1,100 subjects manually.

Additionally, we included in our analysis the most recent sub-cortical sub-segmentations included in FreeSurfer, obtaining measures for the brain-stem substructures (Iglesias et al., 2015), hippocampal substructures (Iglesias et al., 2016), amygdala nuclei (Saygin et al.,

¹⁶Outputs are surfaces instead of being volumetric

2017), and thalamic nuclei (Iglesias et al., 2018).

3.4.2 FreeSurfer QC Results

The results of using Qoala-T to classify subjects can be seen in Figure 3.12:

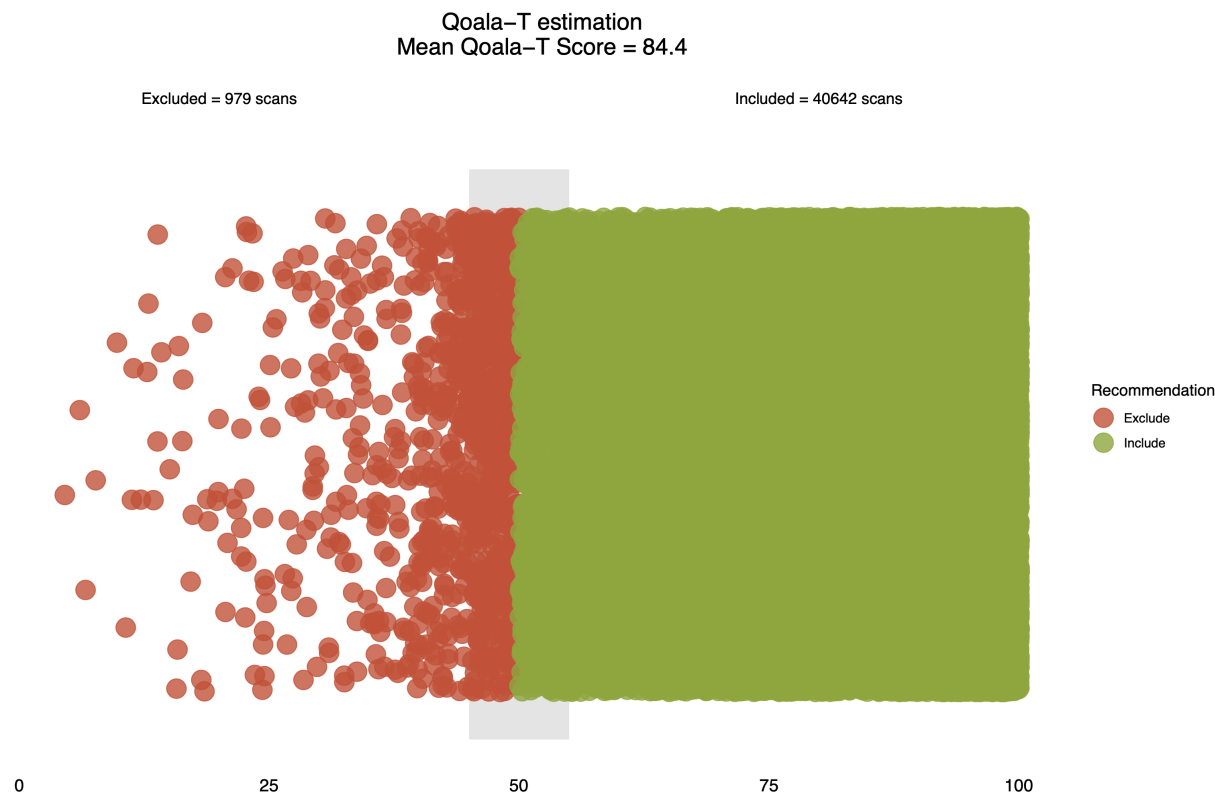


Figure 3.12: Output plot produced by Qoala-T tool with results on all FreeSurfer Datasets. Dots in stroop represent datasets that should be deemed unusable (Qoala-T score $< 50\%$). Dots in red represent usable datasets (Qoala-T scored $> 50\%$). The grey band represents subjects that should be manually checked ($45\% < \text{Qoala-T score} < 55\%$). As can be seen, the vast majority of subjects have a more than acceptable set of results.

If we use the hard threshold of 50%, 979 subjects should be excluded, but we have used the semi-automated approach described above: 558 subjects had a Qoala-T score below 45% and were directly excluded. 1,100 subjects had a Qoala-T score between 45% and 55%. 896 out of those 1,100 presented actual QC problems. Therefore, 1,454 ($558 + 896$) subjects were discarded. A complete account of all FreeSurfer datasets can be seen in Table 3.15.

3.4.2.1 Removing temporal pole data from DKT atlas

After a careful check of the generated FreeSurfer measures, we decided to remove the measures of the temporal pole from some atlases. It is a part of the temporal lobe

difficult to segment, resulting in a massive amount of missing data.

Checking the data of all FreeSurfer right hemisphere-DTK outputs, out of $\sim 40,000$ subjects, only 153 had a volume measure for that structure (meaning that there was not even a row for the temporal pole in the `rh.aparc.DKTatlas.stats` file of the other 40,247 subjects). Out of those 153 subjects:

- 66 had a volume of 0 mm³.
- 47 had a volume of 1 mm³.
- 35 had a volume between 2 and 10 mm³.
- 5 had a volume above 10 mm³ (13, 17, 106, 109, and 1,486 mm³).

The results for the left hemisphere are even worse. No subject out of the $\sim 40,000$ had any data for the temporal pole in the left hemisphere (meaning that there was not even a row for the temporal pole in the `lh.aparc.DKTatlas.stats` file of any subject).

Therefore, we concluded that the metric was not reliable enough to include.

3.4.3 T1 vs T1 + T2 FLAIR

The use of alternative modalities like T2 or FLAIR is recommended in FreeSurfer to refine some parcellations. Nevertheless, there is a not-insignificant number of datasets missing the T2 FLAIR ($\sim 1,350$ datasets out of the $\sim 40,500$ finished datasets). We evaluated the difference in the produced FreeSurfer IDPs with T1 or with T1 and T2 FLAIR in 10 random subjects. Figures 3.13, 3.14, 3.15, 3.16 show these differences, which are especially obvious in the thickness plots.

These differences motivated us to add a new IDP (used as confound in 4.2.4.2), so researchers can control for this factor.

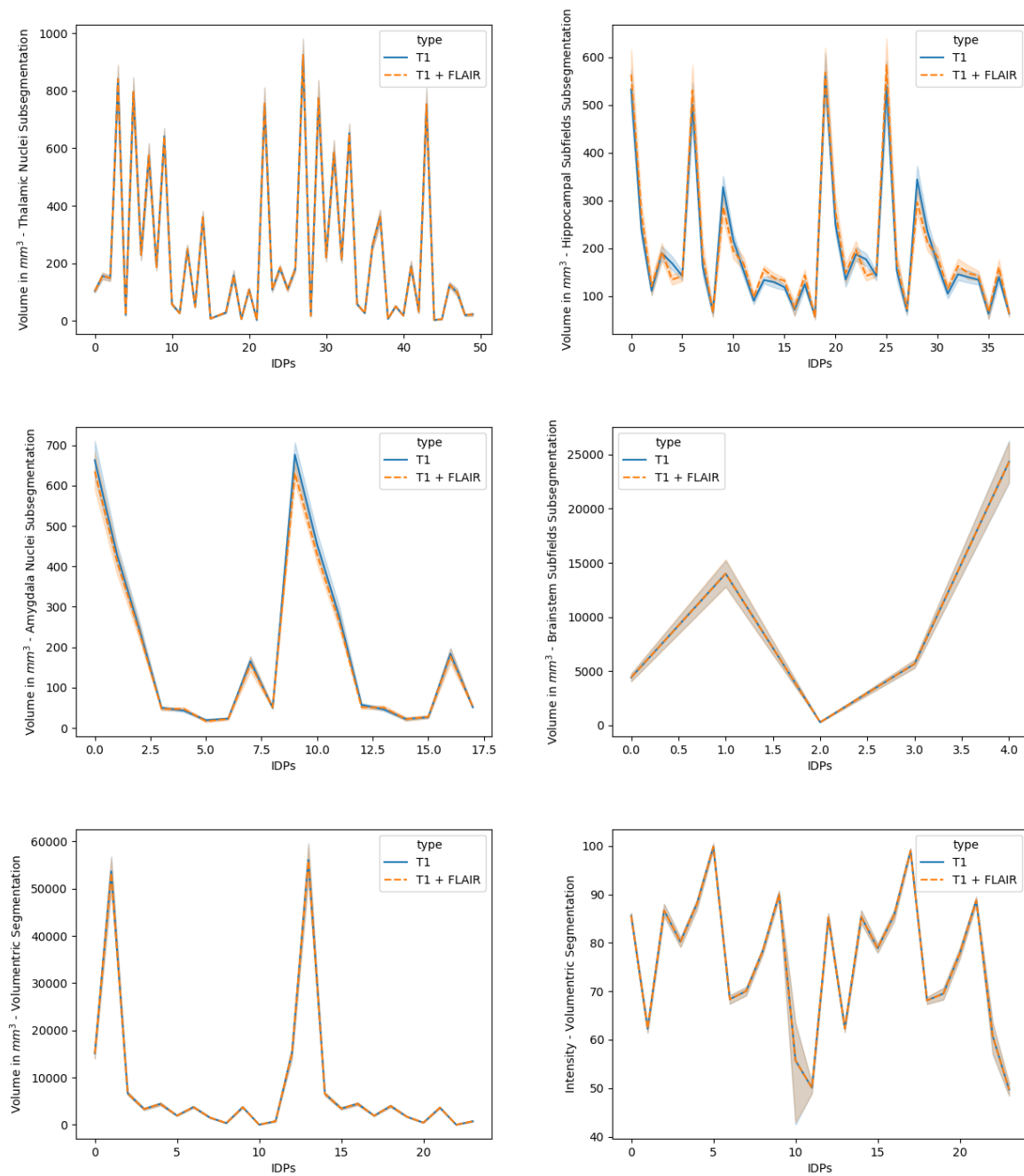


Figure 3.13: Bias in IDPS of volumes IDPs of subcortical segmentations when using T1 vs T1 + T2 FLAIR on FreeSurfer. Lines represent the average of the 10 random subjects, while the confidence intervals show the whole range of values.

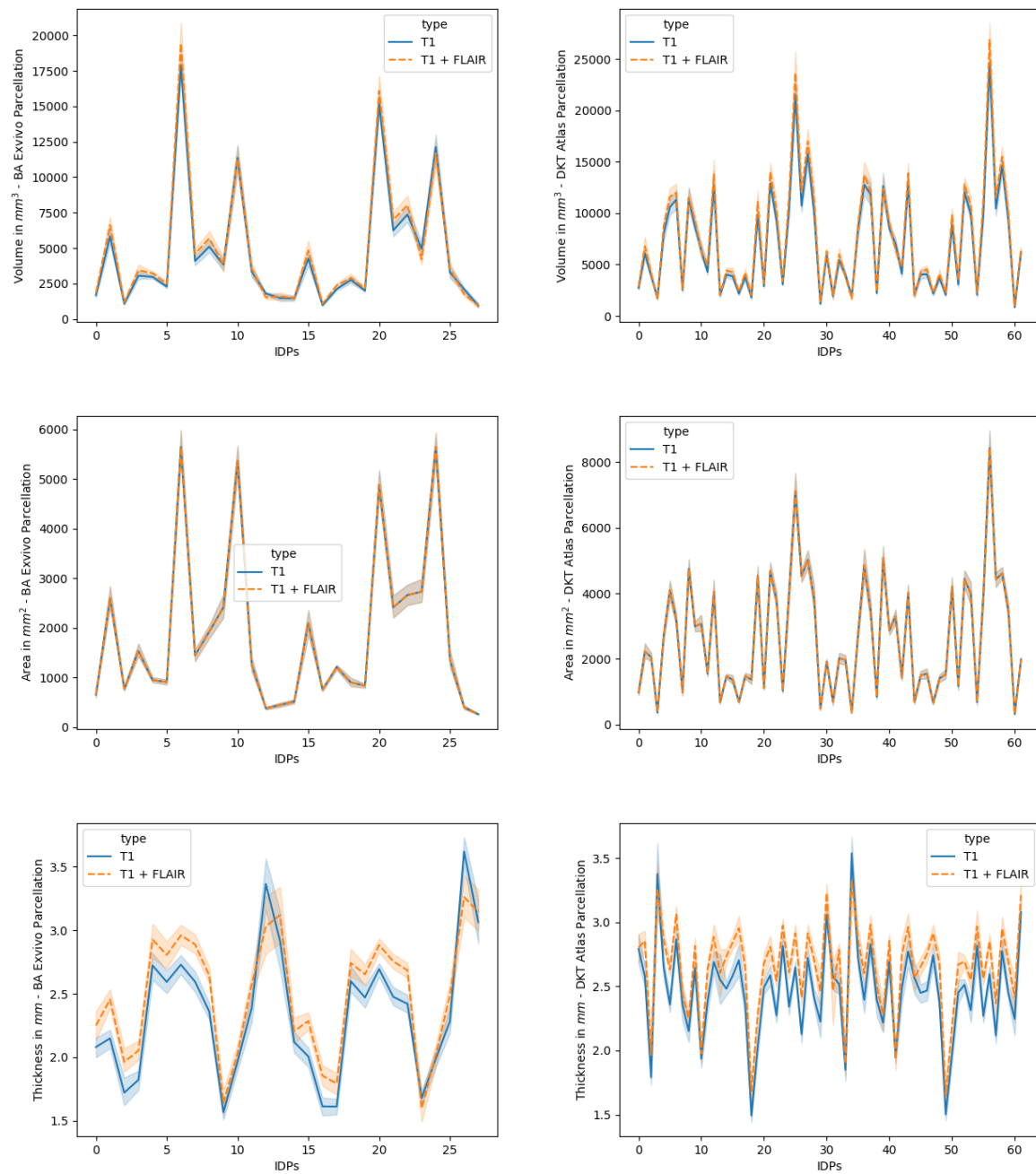


Figure 3.14: Bias in volume, thickness and area IDPS using BA-exvivo and DKT parcellations when using T1 vs T1 + T2 FLAIR on FreeSurfer. Lines represent the average of the 10 random subjects samples, while the confidence intervals show the whole range of values.

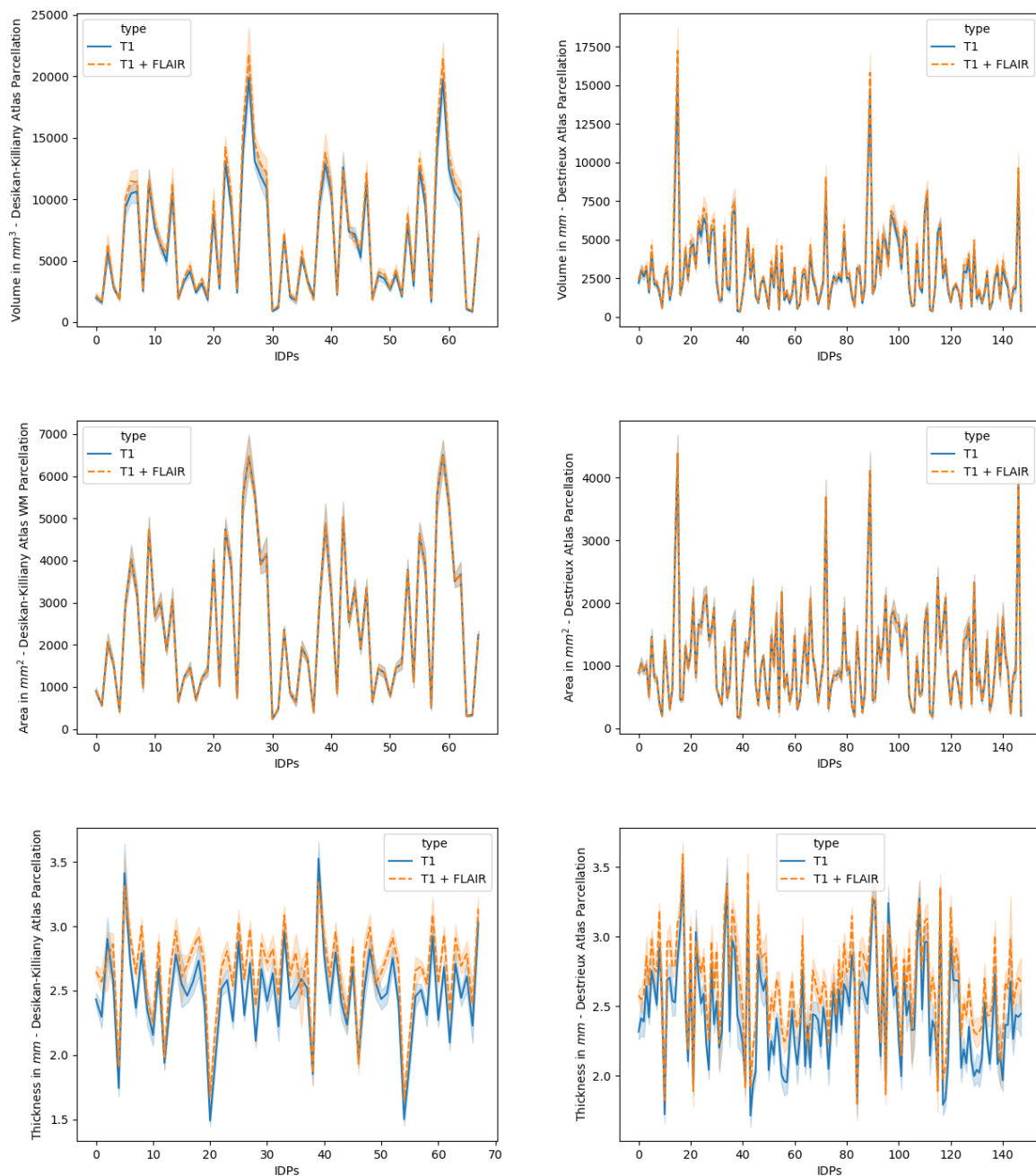


Figure 3.15: Bias in volume, thickness and area IDPS using Desikan-Killiany and Destrieux parcellations when using T1 vs T1 + T2 FLAIR on FreeSurfer. Lines represent the average of the 10 random subjects samples, while the confidence intervals show the whole range of values.

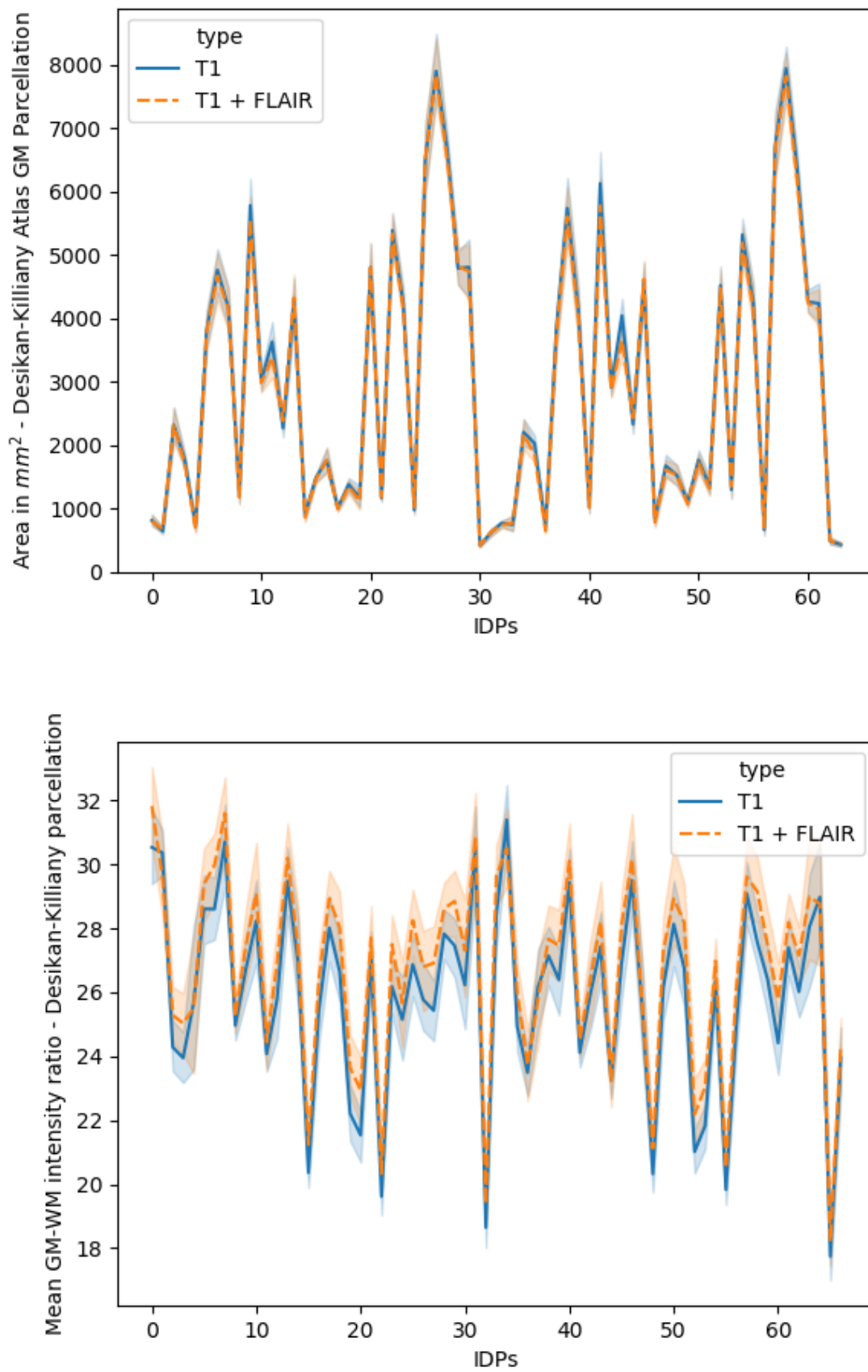


Figure 3.16: Bias in IDPS when using T1 vs T1 + T2 FLAIR on FreeSurfer. Lines represent the average of the ten random subjects samples, while the confidence intervals show the whole range of values.

3.5 Results

Tables with the main results of the QC processes described in this chapter.

		Site				
	QC category	Undetermined*	1	2	3	All
0	Good	0	24,728	9,840	5,111	39,679
1	Unknown issue	0	12	1	4	17
2	Missing modality	450	106	14	0	570
3	Wrong FoV	0	25	0	8	33
4	BM_Noise	0	79	72	40	191
5	BM_Bad	0	137	67	38	242
6	BM_Vent	0	11	7	1	19
7	BM_Fail	0	14	4	3	21
8	BM_Temp	0	444	256	136	836
9	BM_Top	0	311	228	84	623
10	BM_Out	0	106	81	38	225
13	Defacing	0	1	2	1	4
20	Bad_dims	0	2	1	0	3
50	Opted-out	110	3	2	0	115
	Total	560	25,979	10,575	5,464	42,578

Table 3.11: QC classification per site of all datasets after manual re-labelling of the candidates to unusable. All these numbers just refer to T1 images. Note that these numbers are the most updated one and may not match tables in the Appendix.

* We could not determine the site for these subjects, as we did not have their data.

		Site			
	QC category	1	2	3	All
0	Good	95.185%	93.050%	93.540%	93.191%
1	Unknown issue	0.046%	0.009%	0.073%	0.040%
2	Missing Modality	0.408%	0.132%	0%	1.339%
3	Wrong FoV	0.096%	0%	0.146%	0.078%
4	BM_Noise	0.304%	0.681%	0.732%	0.449%
5	BM_Bad	0.527%	0.634%	0.695%	0.568%
6	BM_Vent	0.042%	0.066%	0.018%	0.045%
7	BM_Fail	0.054%	0.038%	0.055%	0.049%
8	BM_Temp	1.709%	2.421%	2.489%	1.963%
9	BM_Top	1.197%	2.156%	1.537%	1.463%
10	BM_Out	0.408%	0.766%	0.695%	0.528%
13	Defacing	0.004%	0.019%	0.018%	0.009%
20	Bad_dims	0.008%	0.009%	0%	0.007%
40	Opted-out	0.012%	0.019%	0%	0.270%

Table 3.12: QC classification percentages per site of all datasets after manual re-labelling of the candidates to unusable.

* We could not determine the site for these subjects, as we did not have their data.

		Processing batch								
	QC category	None	0*	1	2	3	4	5	6	All
0	Good	0	0	5,726	4,209	4,570	6,885	2,070	16,219	39,679
1	Unknown issue	0	0	0	4	1	2	0	10	17
2	Missing mod.	450	10	27	3	68	0	6	6	570
3	Wrong FoV	0	0	2	0	17	3	5	6	33
4	BM_Noise	0	0	6	6	9	26	18	126	191
5	BM_Bad	0	0	47	27	20	28	14	106	242
6	BM_Vent	0	0	0	1	3	4	0	11	19
7	BM_Fail	0	0	0	0	0	0	0	21	21
8	BM_Temp	0	0	5	13	172	184	50	412	836
9	BM_Top	0	0	29	14	117	151	32	280	623
10	BM_Out	0	0	3	5	29	66	18	104	225
13	Defacing	0	0	0	0	0	1	0	3	4
20	Bad_dims	0	0	2	0	0	1	0	0	3
50	Opted-out.	110	0	2	0	0	0	0	3	115
	Total	560	10	5849	4,282	5,006	7,351	2,213	17,307	42,578

Table 3.13: QC classification per processing batch of all datasets after manual re-labelling of the candidates to unusable. All these numbers just refer to T1 images.

* Batch 0 corresponds to the pilot phase described in Section 1.3.3.

		Processing batch						
	QC category	1	2	3	4	5	6	All
0	Good	97.897%	98.295%	91.290%	93.661%	93.538%	93.714%	93.191%
1	Unknown issue	0%	0.093%	0.020%	0.027%	0%	0.058%	0.040%
2	Missing mod.	0.462%	0.070%	1.358%	0%	0.271%	0.035%	1.339%
3	Wrong FoV	0.034%	0%	0.340%	0.041%	0.226%	0.035%	0.078%
4	BM_Noise	0.103%	0.140%	0.180%	0.354%	0.813%	0.728%	0.449%
5	BM_Bad	0.804%	0.631%	0.400%	0.381%	0.633%	0.612%	0.568%
6	BM_Vent	0%	0.023%	0.060%	0.054%	0%	0.064%	0.045%
7	BM_Fail	0%	0%	0%	0%	0%	0.121%	0.049%
8	BM_Temp	0.085%	0.304%	3.436%	2.503%	2.259%	2.381%	1.963%
9	BM_Top	0.496%	0.327%	2.337%	2.054%	1.446%	1.618%	1.463%
10	BM_Out	0.051%	0.117%	0.579%	0.898%	0.813%	0.601%	0.528%
13	Defacing	0%	0%	0%	0.014%	0%	0.017%	0.009%
20	Bad_dims	0.034%	0%	0%	0.014%	0%	0%	0.007%
50	Opted-out.	0.034%	0%	0%	0%	0%	0.017%	0.270%

Table 3.14: QC classification percentages per processing batch of all datasets after manual re-labelling of the candidates to unusable. All these numbers just refer to T1 images.

Class	Number of Datasets
Available & complete FS datasets	40,014*
Missing T1	570
Subject opted-out	115
FS failed / unfinished**	153
Discarded after previous T1 QC***	272
Discarded after Qoala-T QC****	1,454
Total	42,578

Table 3.15: QC classification per FreeSurfer dataset.

* The number of available datasets does not match the number in the UK Biobank Data Showcase (40,060) because we are not including repeated scans here.

** 153 subjects (0.4%) of the total FreeSurfer runs failed (Error state) or were manually stopped after having run for more than 40 hours

*** Subjects from QC categories 3, 5, 6, 7, 13, and 20 did not go through FS processing, as FreeSurfer developers discourage trusting results with those problems.

**** 558 subjects had a Qoala-T score below 45%. 1,100 subjects had a Qoala-T score between 45% and 55%. 896 out of those 1,100 presented actual QC problems. Therefore, 1,454 (558 + 896) subjects were discarded.

3.6 Discussion

The considerable number of subjects makes the need for an at least partially automated tool evident, including the detection of inadequate datasets or failures in processing. For this reason, the development of an adequate automated QC tool is an absolute necessity, and we have presented here an approach for QC of T1 data that includes this.

The study of possible drifts in imaging data over time will be essential and will be the subject of the next chapter. We will investigate the possible effects of minor scanner software upgrades and minor protocol changes, as well as inter-site imaging consistency. We will also investigate the extent to which simple adjustment for such drifts may be effective.

A future long-term challenge will be data analysis and population modelling. The application of new unsupervised learning methods for finding new data-driven IDPs or QC metrics is potentially interesting (Duff et al., 2015). This could also evolve to consider health outcomes as noisy/weak labels (Frénay and Verleysen, 2014) in such a way that the learning process considers their uncertainty and reliability.

The development of the QC tool allowed us to find problematic steps in the processing pipeline (i.e., non-linear registration to the MNI template). Even though the number of subjects with this problem is not large ($103/5816 = 1.8\%$, compared to the 2–13% failure

rate for the acquisitions), and we can detect them fairly consistently, improving this step will be a future line of work. Possible improvements may involve a combination of the following:

- Generating a study-specific T1 template and registering subject images to it instead of to the MNI template.
- Improving brain extraction and using the resulting brain mask to guide registration. Possibilities include running FNIRT with much lower degrees of freedom to initialise brain extraction or developing a classifier that works directly on each voxel to classify them as brain/non-brain. See the Appendix for more details on the brain extraction method we have used to date.
- Combining T1+dMRI (and possibly more modalities) in the registration in a more integrated manner; for example, taking advantage of the richer signal within white matter available from the diffusion imaging. A new tool called MMORF ([Lange et al., 2020](#)) could be helpful for this.

A possible way to improve QC accuracy would be in the development of better features to drive the classification. These new features may be created by finding new heuristics that better describe the problems or by using unsupervised machine learning techniques to model the data, e.g., using ICA-like techniques ([Hyvärinen, 1999](#); [Beckmann and Smith, 2004](#)).

Also, we need the development of metrics to describe the discrepancy between the T1 image (for a given subject) and each of the other modalities (for that same subject), after linear alignment of the other modalities to the T1. Therefore, we are planning to extend the automated QC system to reliably find problems in the acquisition and processing steps for all modalities (not just T1). Again, ICA-like techniques may be useful here, as well as methods based on the Hidden Markov Model ([Baker et al., 2014](#); [Vidaurre et al., 2016](#)) to interrogate the network dynamics.

Finally, an important limitation of this work has been the subjectivity of the manual labelling of the QC candidates as unusable, described in Section 3.3. There has been only one rater evaluating these candidates, and a more comprehensive multi-rater scheme would have been more reliable. As an example, the scheme described in [Klapwijk et al. \(2019\)](#), dividing the rating tasks between several raters and having some overlap between

those tasks to allow the evaluation of the agreement between raters and the reliability of each rater may be better, but to date, this has not been logistically possible.

As an example of how this subjectivity may have affected the evaluations, Table 3.16 shows the number of subjects with a different QC evaluation between their first imaging visit second imaging visit, two years later.

#	Subjects	QC in 1st visit	QC in 2nd visit
1	1	0 - Good	5 - BM_Bad
2	1	2 - Missing Mod.	0 - Good
3	1	4 - BM_Noise	0 - Good
4	1	9 - BM_Top	10 - BM_Out
5	1	9 - BM_Top	4 - BM_Noise
6	1	9 - BM_Top	8 - BM_Temp
7	2	5 - BM_Bad	0 - Good
8	2	8 - BM_Temp	10 - BM_Out
9	4	0 - Good	10 - BM_Out
10	5	0 - Good	4 - BM_Noise
11	5	3 - FOV	0 - Good
12	6	10 - BM_Out	0 - Good
13	9	0 - Good	9 - BM_Top
14	30	9 - BM_Top	0 - Good
15	31	8 - BM_Temp	0 - Good
16	33	0 - Good	8 - BM_Temp

Table 3.16: Number of subjects with every discrepancy in the QC evaluations between the first and the second imaging visit.

Even though most of the discrepancies are plausible (a subject may move in the second visit and not in the first one and *viceversa*, or the registration may have failed in just one of the visits or failed differently in each visit), circumstances like the 7th line, where two subjects were labelled with category 5 (BM_Bad: structurally atypical) in the first visit but, labelled as “Good” in the second visit are rather unlikely. Conversely, the first line when one subject was labelled as “Good” in the first visit and “structurally atypical” in the second visit, may be plausible due to a pathology happening in the two-year interval between those visits.

Chapter 4

Confound modelling

4.1 Introduction: Dealing with confounds

Dealing with information coming from UK Biobank without careful treatment of possible confounding factors is problematic for several reasons: spurious associations can be induced between pairs of otherwise independent variables if the unconfounding is not carried out correctly (e.g., if the confounds were not demeaned first); the significance of real associations can be biased, and therefore their interpretability affected; confounding factors can be erroneously estimated, and hence regressing them out of the data can be ineffective ([Westfall and Yarkoni, 2016](#)); finally, there can be instances of Berkson’s paradox, where a variable is incorrectly treated as representing a causal confounding process. While confounds are a potential problem for datasets of any size, the large N setting is particularly challenging due to even very small confounding effects causing misleading results. For further discussion on the importance of confounds in imaging research, see ([Smith and Nichols, 2018](#)).

The consideration of a variable as a confound depends heavily on the context; for example, age can be a confounding factor in some studies, while being a variable of interest in others. Another example is sex, which correlates with potential confounds (such as head size), and which can also influence variables of interest in complex ways (e.g., the trajectory of bone density with ageing). Complicated confounds such as sex may force the researcher to carry out separate association analyses for the different sexes (as opposed to merely regressing out the sex variable). Hence, this chapter is not attempting to address issues raised where variables can contain both signal of interest and confounding factors (given that the answer will depend on the context of the biological question being asked); instead, here we focus on investigating the extent to which potential confounding factors explain variance in the UKB brain imaging data, and also the effects of unconfounding on correlations between Imaging Derived Phenotypes (IDPs¹) ([Alfaro-Almagro et al., 2018](#)) and non-

¹IDPs (imaging-derived phenotypes) are individual measures of brain structure and function, such as

imaging variables (non-IDPs or nIDPs).

The question of how to deal with confounds after they have been identified has been investigated in previous literature. Many studies ((Dukart et al., 2011), (Kostro et al., 2014), (Rao et al., 2017)) either regress out the confounds from the data or use them as additional regressors in their (e.g., multiple regression) analyses. Alternative methods have been suggested, like restriction (i.e. limiting the study to subjects with a certain feature, as shown in (Zarnani et al., 2019) with a cohort study centred on males with the same age and nationality), matching subjects for a certain confound (e.g. sex and age (Rosenbaum and Rubin, 1983), which can be done either a priori, before the acquisition, or a posteriori, by selecting certain subjects for the analysis), stratification (i.e. dividing the data into different hierarchical levels, according to the different features to unconfound), or representing the data in a way that is insensitive to confounding factors (Glastonbury et al., 2018). Due to the large number of confounds that we are dealing with in the UKB data, our general approach is to regress out the confounds from the data (and model, where appropriate), although many of the results presented below should be relevant in the context of other unconfounding strategies. For further discussion, see (Jager et al., 2008) and (Snoek et al., 2019).

Here we consider confounds related to the acquisition process and scanner configuration, subject-specific biometric variables, motion confounds, acquisition date/time confounds, and table position. We model these confounds in several ways and explore how the data of the first $\sim 40,000$ subjects is affected by the modelling. We compare our set of confounds with a more “traditional” smaller set of confounds. Finally, we outline a set of recommendations on how to use this information when running studies using UKB brain imaging data.

4.2 Methods

4.2.1 Linear model unconfounding & Variance measures

Our unconfounding is performed simultaneously in this study (i.e., in a single regression-based unconfounding using all confound variables together) with a linear model. Let the

the volume of specific brain structures or the strength of connectivity between pairs of brain regions. Non-IDPs (or nIDPs, non-imaging-derived phenotypes) are other phenotypes not derived from the brain imaging, such as body weight, specific cognitive test scores, or disease diagnoses

N -vector Y be one variable of interest, and X the N -by- P matrix of all confounds, then the confound-adjusted variable is the residuals from the linear model fit of Y to X , i.e. $Y - XX^-Y$, where X^- is the Moore-Penrose pseudo-inverse of X . There is no intercept term in the model, as all confound variables are demeaned.

When measuring variance explained by a subset of confounds, say matrix X_1 , where all confounds are $X = [X_1 \ X_2]$, we define percent variance explained by X_1

$$\%VE = 100 \times \hat{Y}_1' \hat{Y}_1 / Y'Y,$$

and percent unique variance explained, the additional variance explained by X_1 not already explained by X_2 ,

$$\%UVE = 100 \times (\hat{Y}'\hat{Y} - \hat{Y}_2'\hat{Y}_2) / Y'Y,$$

where $\hat{Y}_2 = X_2 X_2^- Y$ is the prediction using X_2 alone, and $\hat{Y} = X X^- Y$ is the prediction using all confounds.

We did not seek to obtain fully unbiased estimates of the (potentially biased, although minimally so due to the very large N) $\%VE$ by each type of confound, e.g., through cross-validation. We simply use the all-data-estimated $\%VE$ to rank and prioritise the many different possible confounds. Note that, even if cross-validation were used to obtain non-circular estimates of $\%VE$, an arbitrary threshold would still have to be applied to select included confounds.

4.2.2 Data description

This chapter used imaging and non-imaging data from UKB, accessed under data access application 8107. The majority of the work reported in this chapter was carried out using the 22,000 subjects May 2019 data release, and then the final results were updated using the 40,000 subject January 2020 data release.

The January 2020 data release includes 41,985 datasets from the first three UKB sites: 25,962 from Stockport (Site 1), 10,560 from Newcastle (Site 2), and 5,463 from Reading (Site 3). Scanning at the fourth site in Bristol began in early 2020. Additionally, 1,587 subjects (1,117 from Site 1 and 470 from Site 2) have been scanned a second time (with a mean of 2 years difference between the two scans). Although this repeat-scan data has been released, it was not used here, as the study of the longitudinal data is outside the

scope of the present work.

After removing datasets that were deemed unusable by our QC procedure (for example, due to having incomplete brain coverage, or having very severe MRI artefacts) ([Alfaro-Almagro et al., 2018](#)), the number of subjects’ datasets that were analysed in this work is 39,694 (21,005 females).

4.2.3 Overview of confound types

Figure 4.1 shows how each group of confounds is related to each other group by the percentage of variance that one explains in the other. Each row/column relates to a single confound group (e.g., imaging site); a given confound “group” might be implemented in the unconfounding as multiple confound variables (e.g., a separate binary indicator variable for each imaging site). This matrix of % variance explained is fairly symmetric, but not exactly (because different groups, in general, contain different numbers of variables).

In order to help describe the confounds we have identified, we arranged the confound groups into six different families: general subject-specific features, scanner/acquisition protocol/processing parameters, head motion, scanner table position, non-linear and “crossed terms” (interactions between different confounds), and acquisition date and time.

We now describe the first four confound families in more detail (noting that these four are also the “source data” for the last two families).

4.2.4 Description of basic confound variables (80 variables)

There are four families of basic confounds: Subject-specific confounds, acquisition/processing parameters, head motion and table position.

4.2.4.1 Subject-specific confounds (4 variables)

The basic confounds in this family are age, sex, the product of age and sex, and head size scaling. The first two are taken from UKB non-imaging variables, while the latter was calculated with SIENAX ([Smith et al., 2002](#)) as part of the T1 processing pipeline 2.2.2. It is defined as a ratio that shows the volumetric scaling from the T1 head image to MNI standard atlas. This set of confounds is commonly used in many brain imaging studies. A discussion about the type of studies for which these confounds may be useful can be read in ([Barnes et al., 2010](#)).

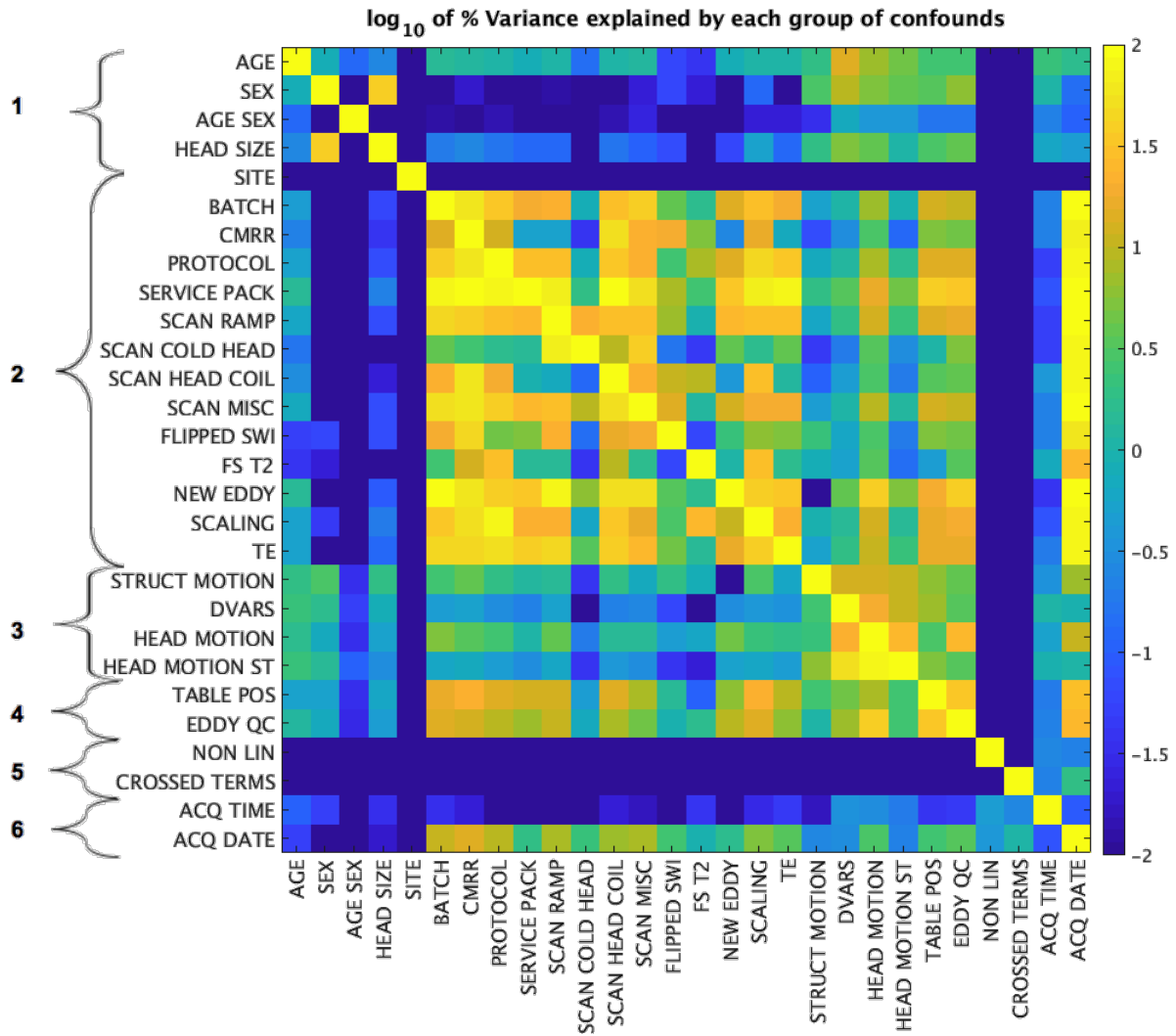


Figure 4.1: Matrix showing the percentage of variance of each group of confounds explained by each other group. Each row and column represents one group of confounds. These groups can be organised into families: 1: Subject-specific confounds; 2: Scanner acquisition protocol processing parameters; 3: Head motion confounds; 4: Table-position-related confounds; 5: Non-linearities and crossed terms; 6: Date/time-related confounds. The site group was forced to be independent of the other confound groups, as described in Section 4.2.5.1. This means that for later analysis, the site group is only explaining variance not already explained by other variables. Nonlinearities and cross-terms are forced by definition to be orthogonal to linear terms. Independence from all other confound groups was also forced for acquisition time and date, but there may be some random correlations with date because of the smoothing described in Section 4.2.5.4. An interactive version of this figure showing the actual values in each element of the matrix can be found in [\[LINK\]](#).

4.2.4.2 Scanner / acquisition protocol / processing parameters (20 variables)

Any differences in scanner hardware, configuration, acquisition protocol or processing parameters can affect the imaging data, and should therefore be modelled as confounds ((Focke et al., 2011), (Chen et al., 2014), (Keenan et al., 2019)). UKB is using identical scanner hardware and software in all sites (3T Siemens Skyra, 32-channel Siemens head RF coil, software VD13), but having the acquisition site as a confound (SITE) may be important, as there might be subtle differences (for example, differences in different RF coils even of the same model).

Scanner servicing and minor changes in acquisition parameters may also affect the data, and may therefore need to be considered in the confound modelling. Therefore, we investigated the CMRR² multiband software version (8 versions, with minor version changes between these), scanner Service Pack software version (2 versions), and hardware/servicing events in the scanner. Previous studies show that such changes may bias IDPs ((Krueger et al., 2012), (Noble et al., 2017)). These hardware events are summarised in four different variables:

- 1 B0 field ramp-down/up events (SCAN RAMP): Four events in Site 1.
- 2 Head Coil replacements (SCAN HEAD COIL): Three events in Site 1 and one in Site 2.
- 3 Cold Head replacement: (SCAN COLD HEAD): One event in Site 1, two events in Site 2.
- 4 Miscellaneous events (SCAN MISC): Three events in Site 1, two events in Site 2.

It has also recently become clear that there have been slowly-changing heating-related effects in the extent of eddy currents in the diffusion MRI (dMRI) data. This effect is now regularly checked for, and the scanner recalibrated when appropriate. However, affected datasets needed to have a more robust version of eddy current correction applied (primarily by increasing the search space for eddy currents when using the Eddy preprocessing tool). A new confound variable reflecting this effect has been created (NEW EDDY).

We also considered minor changes in the acquisition protocol that, in principle, should not affect the data (6 phases described in Section A.3 of the Appendix), and a temporary unintended protocol change³ in the Susceptibility Weighted Images (SWI) acquisition that

²CMRR: Centre for Magnetic Resonance Research (www.cmrr.umn.edu)

³A change in phase encoding direction - not a minor thing for some acquisitions, but quite minor for

affected 3,355 datasets (Variables PROTOCOL and FLIPPED SWI).

A few minor protocol parameter changes have been made in error at the imaging sites for a small subset of subjects, in the process of distributing the protocol across sites (fMRI echo times of 42.4 vs 39ms, and overall global intensity scaling of images). Therefore, we included six variables (SCALING), one per modality, as confounds, and also the echo times (TE) for resting-state fMRI (rfMRI) and task fMRI (tfMRI). As seen below, none of these has a large effect on derived IDPs.

Finally, the processing of the $\sim 40,000$ subjects, as described in (Alfaro-Almagro et al., 2018) was performed in six separate batches. Both the operating system and the processing pipeline software used were kept almost unchanged for each batch, but there were differences in the processing hardware used across batches, so we wanted to make sure that these differences did not affect the data in any way (BATCH).

Due to the absence of T2 FLAIR image in some subjects, FreeSurfer was not processed in the same way for every dataset; 1,301 datasets were processed using just the T1w images, while the vast majority of datasets (38,173) used both T1 and T2 FLAIR in the FreeSurfer processing. A confound variable (FS T2) describes whether the T2 FLAIR was used.

The numbers of subjects for all these parameters, as well as some more details, are listed in Section A.3 in the Appendix.

4.2.4.3 Head Motion (48 variables)

Four confound groups capture subject head motion during acquisition, both in the 4D modalities (task fMRI, resting fMRI and dMRI) and structural modalities (e.g. T1w). The importance of head motion as a confound has been known for a long time (Friston et al., 1996). As noted in (Greve et al., 2013), some studies measure and compensate for motion prospectively (i.e. before the analysis), while many others estimate it in a post-acquisition phase, being a standard step in most MRI processing packages. Most studies use these motion estimates as covariates in a GLM analysis (Johnstone et al., 2006), as a criterion to remove certain volumes from 4D images (Power et al., 2014) or to regress them out from the global signal (Murphy and Fox, 2017). For interesting discussions on the different methods of accounting for motion, see (Satterthwaite et al., 2013) and

(Murphy et al., 2013).

In our study, one approach we took to modelling head motion was to estimate it from the structural images. Estimations of motion-induced artefacts in structural images are related to motion estimates in temporal modalities in complex ways (Savalia et al., 2017). However, the existence of a strong relationship of some of these estimates with valid non-artefactual structural measures of the brain (such as grey matter volume) is a problem (Gilmore et al., 2019).

We estimated the structural motion by fitting a cross-validated linear regression where the dependent variable was a manually evaluated QC measure of motion in 871 T1w images, and the independent variables were a set of features that are related to structural motion and QC (e.g. smoothness estimates in X, Y, and Z (Flitney and Jenkinson, 2000), average Euler number of the FreeSurfer surfaces (Rosen et al., 2018), Qoala-T quality metric of FreeSurfer output (Klapwijk et al., 2019), etc.). This resulted in one variable (STRUCT MOTION) summarising the motion in the structural acquisition, to be included as a confound. How this metric was calculated was described in detail in Section 3.3.3; in terms of automatically predicting the expert-judged motion QC score (on a scale of 1-4), the trained predictor had an MSE of 0.14 (predicting data of range 0:1) in the left-out set.

Many studies of head motion focus on “temporal” imaging modalities (fMRI, dMRI, etc.). Hence, we obtained the motion estimates from FSL’s FEAT (Woolrich et al., 2001) and Eddy ((Andersson et al., 2016), (Andersson et al., 2017)), and estimated the mean, median and 90th percentile over time of the absolute and relative motion (averaged across space) in the task fMRI, resting fMRI, and dMRI. We also included (as a confound) the number of slices that Eddy estimated to be outliers in the dMRI data (because of significant signal dropout which is mostly due to motion). This resulted in 19 confound variables (HEAD MOTION). A second approach has been to calculate the same quantile summaries (mean, median and 90th percentile) of the motion over space and time calculated from FSL’s FEAT motion estimation matrices from resting fMRI in a similar way as described in (Satterthwaite et al., 2013) (HEAD MOTION ST, 10 variables). These might capture additional useful motion-related confound information given that the amount of motion varies across both space and time in general.

Finally, we calculated the mean, median and 90th percentile over time of S-var and D-var normalised by A-var (variants of DVARs ([Afyouni and Nichols, 2018](#))) from both the original resting fMRI and the resting fMRI after removal of noise components using FIX (see ([Griffanti et al., 2014](#)), ([Griffanti et al., 2017](#))) (DVARs, 18 variables).

4.2.4.4 Table position (8 variables)

Early in the project, we detected that the head position and the scanner-table position (meaning, in effect, the position of the RF head coil within the scanner, relative to isocenter) were correlated with several QC metrics and IDPs. The most important factor is the location of the coil/head in the scanner in the direction that the scanner table moves in and out, although the precise cause of this effect has not been established. Therefore, we included these parameters as possible confounds.

The first set of confounds relates to positions of the head and RF coil relative to the scanner (i.e., relative to the isocentre of the scanner). This set consists of the Z-position of the coil (more specifically, the scanner table on which the coil is positioned) within the scanner, as read from the DICOM headers; the X and Z coordinates of the Centre Of Gravity of the T1w brain mask; and the Y position of the most posterior part of the same brain mask (TABLEPOS, 4 variables).

We also noted how measures from QUAD (a recent QC tool for dMRI QC ([Bastiani et al., 2019](#))) were highly non-linearly correlated with the table position (See Figure 4.2. For this reason, four metrics from QUAD were included as confounds (EDDY QC, 4 variables):

- 1 Standard deviation of X, Y, and Z volume-wise components of the estimated eddy currents' linear field (columns 7, 8, and 9 from the eddy parameters output file); these should primarily reflect eddy currents
- 2 Standard deviation of Y volume-wise component of the translations due to the subject's head motion (column 2 from the eddy parameters output file).

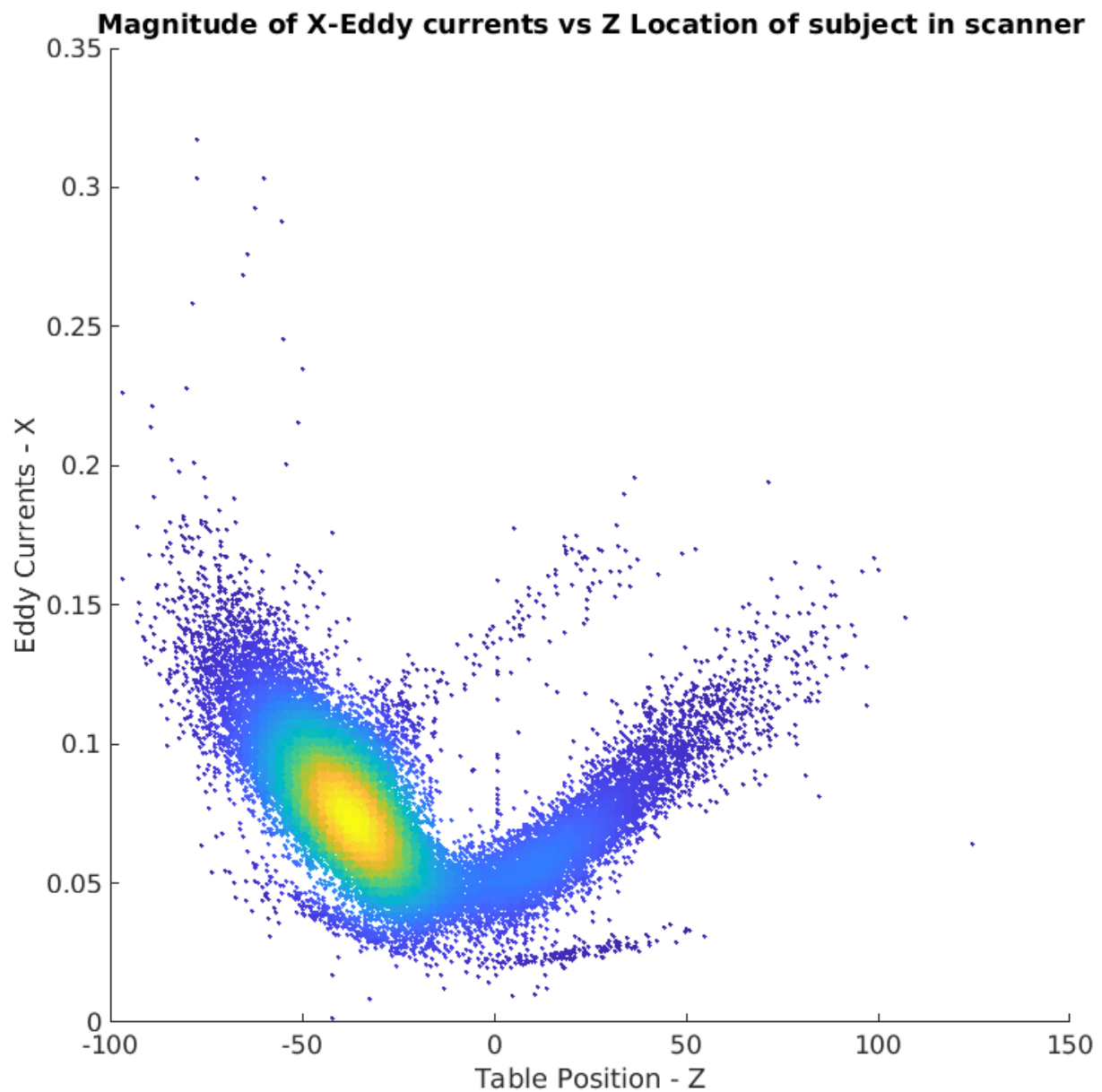


Figure 4.2: Complex relationship between Table Position (Z) and the Eddy Currents (X). All subjects (40,000) were used for this plot.

4.2.5 Processing of confound regressors (602 variables)

Table 4.1 contains a list of all confound groups, the availability of the confound in UKB (see Section 4.4.6), the number of variables in each group, the number of variables after expanding / processing those confounds and an indication of the group being either qualitative or quantitative.

Conf. family	Confound group	UKB ID	# Confs	# Processed	Processing	Section
SUBJECT	AGE	53 - 33	1	3	Quantitative	4.2.4.1
	SEX	31	1	3	Qualitative	4.2.4.2
	AGE SEX	N/A	1	3	Quantitative	4.2.4.1
	HEAD SIZE	25000	1	3	Quantitative	4.2.4.1
ACQUISITION	SITE	54	1	2	Qualitative	4.2.4.2
	BATCH	TBD	1	7	Qualitative	4.2.4.2
	CMRR	TBD	1	8	Qualitative	4.2.4.2
	PROTOCOL	TBD	1	5	Qualitative	4.2.4.2
	SERVICE PACK	TBD	1	1	Qualitative	4.2.4.2
	SCAN RAMP	TBD	1	4	Qualitative	4.2.4.2
	SCAN COLD HEAD	TBD	1	2	Qualitative	4.2.4.2
	SCAN HEAD COIL	TBD	1	4	Qualitative	4.2.4.2
	SCAN MISC	TBD	1	5	Qualitative	4.2.4.2
	FLIPPED SWI	TBD	1	3	Qualitative	4.2.4.2
	FS T2	26500	1	3	Qualitative	4.2.4.2
	NEW EDDY	25921	1	1	Qualitative	4.2.4.2
	SCALING T1	25925	1	1	Qualitative	4.2.4.2
	SCALING FLAIR	25926	1	2	Qualitative	4.2.4.2
	SCALING SWI	25927	1	2	Qualitative	4.2.4.2
	SCALING dMRI	25928	1	2	Qualitative	4.2.4.2
	SCALING rfMRI	25929	1	1	Qualitative	4.2.4.2
	SCALING tfMRI	25930	1	1	Qualitative	4.2.4.2
	TE rfMRI	25923	1	3	Quantitative	4.2.4.2
	TE tfMRI	25924	1	3	Quantitative	4.2.4.2
MOTION	STRUCT MOTION	TBD	1	3	Quantitative	4.2.4.3
	DVARS	TBD	18	54	Quantitative	4.2.4.3
	HEAD MOTION	TBD	19	57	Quantitative	4.2.4.3
	HEAD MOTION ST	TBD	10	30	Quantitative	4.2.4.3
TABLE	SCAN POSITION X	25756	1	3	Quantitative	4.2.4.4
	SCAN POSITION Y	25757	1	3	Quantitative	4.2.4.4
	SCAN POSITION Z	25758	1	3	Quantitative	4.2.4.4
	SCAN TABLE POS.	25759	1	3	Quantitative	4.2.4.4
	EDDY QC	TBD	4	12	Quantitative	4.2.4.4
NON LINEAR	NON LINEAR	N/A	549	158	Non-linear	4.2.5.2
X-ED TERMS	CROSSED TERMS	N/A	79,003	84	Crossed-term	4.2.5.3
TIME	ACQ TIME	53	1	61	PCA	4.2.5.4
	ACQ DATE	53	1	59	PCA	4.2.5.4

Table 4.1: Size of confound groups before and after processing

TBD - To Be Decided: Source variables to be sent to UKB to be made available for researchers shortly.

N/A - Not Applicable: Non-source variables that will not be sent to UKB.

4.2.5.1 Basic confounds (80 variables, expanded to 240)

The 80 basic confounds were partially processed to account for interactions between site and other confounds, due to site potentially being one of the most important confounds

in any multi-site study⁴. For example, we might expect head-size to act as a confound slightly differently in different sites, so we create separate head size confounds for the different sites. This processing ends up expanding the initial set of confounds to a total of 240 confounds⁵. The processing steps are as follows:

- 1 *Separation by site, demediation and outlier removal of quantitative confounds*: All variables belonging to a given confound group (e.g. X, Y, Z brain centre of gravity + table position) are demediated and normalised (using the median absolute deviation) globally (i.e. across all sites). Then, each variable is replicated as many times as sites we are dealing with, but each copy only retains values for subjects which were scanned at the corresponding site. The variable value for all other subjects is replaced with 0s. Then, all outliers⁶ and all missing values in each copy are replaced with the median value for the site. Finally, each copy is normalised separately to have zero mean and unit standard deviation.

Pseudo-code describing this step is included in the Appendix, Section A.4.

- 2 *Separation by site, binarisation and normalisation of qualitative confounds*: For each categorical confound (such as CMRR version or Service Pack), site-independence is enforced by first binarising and expanding the original confound (a confound with n possible categorical values will be replaced by $n-1$ binary indicator variables) and duplicating the resulting set by site, provided the site has more than one different value for the confound. Finally, and as in the previous case, each resulting variable is de-mediated by site.

Pseudo-code describing this step is included in the Appendix, Section A.4.

⁴All processing code used in this study is available online at www.fmrib.ox.ac.uk/ukbiobank/confounds/

⁵Non-confound IDPs are used in this work both as part of generating confounds (see time and date drift confounds below) and in judging the effects of unconfounding on IDPs. To reduce the effect of potential outliers and improve the accuracy of associations, we applied rank-based inverse Gaussian transformation (Quantile Normalisation, or QN) on all IDPs and nIDPs to impose Gaussianity (Bolstad et al., 2003) and resulting in variables being zero mean, before using them in any work described here

⁶For any given confound, we define outliers thus: First, we subtract the median value from all subjects' values. We then compute the median-absolute-deviation (across all subjects) and multiply this MAD by 1.48 (so that it is equal to the standard deviation if the data had been Gaussian). We then normalise all values by dividing them by this scaled MAD. Finally, we define values as outliers if their magnitude is greater than 8

4.2.5.2 Non-linear transformations (158 variables)

The existence of non-linear effects in confounding variables has been discussed in previous studies. For example, (Barnes et al., 2010) shows that ratio-based or linear unconfounding for head size is not sufficient for many studies, and (Smith and Nichols, 2018) suggest that *‘adding transformed versions of confounding variables will allow more than just linear effects of confounds to be captured’*.

We decided to perform three different non-linear transformations to all quantitative confound variables to capture possible non-linear effects. These transformations would be:

- 1 Squaring the centred confound.
- 2 Quantile Normalisation (QN) of the confound, forcing it to have a Gaussian distribution (Bolstad et al., 2003).
- 3 Squaring the Quantile Normalisation.

These transformations were applied to the 183 quantitative confound variables, resulting in a new confound group of 549 non-linear confounds. Nevertheless, not all these confounds are equally important, and increasing the number of confounding variables is both inconvenient in computational terms but can also unnecessarily remove too many degrees-of-freedom from the data (confounds that are just random noise explain a certain amount of variance of the IDPs, as described in Section 4.2.6). We estimated the amount of unique variance explained (in the IDPs) in order to decide which non-linear confounds to keep, and ended up retaining 158 confounds (NON-LINEAR). The criteria that we used (described in detail in Section 4.3.1) is that each non-linear confound must pass at least one of two tests, to be included in the confound list:

- 1 Average (across IDPs) of the % VE (percent variance explained) by the non-linear confound must be higher than the 95th percentile of all average % VEs (across IDPs).
- 2 Maximum (across IDPs) % VE by the confound must be higher than the maximum of 0.75 % VE and the 99.9th percentile of all VEs (across IDPs).

The thresholds used here and in the next section are by necessity empirically “expert” determined (and arguably arbitrary), but can, of course, be changed if researchers use our unconfounding code and wish to make unconfounding more or less aggressively conservative.

4.2.5.3 Crossed terms (Confound interaction: $A * B$) (84 variables)

The confounds may interact with each other in a non-additive way. Thus, products of pairs of confounds should be considered. Pairwise products of the 398 confound variables ($240 + 158$) would produce new confound variables (79,003). Combinations of confounds from different sites will be useless (the product of those will be 0), so we end up evaluating 26,674 crossed-terms. As previously mentioned, not all these confounds are useful, so we used similar criteria as for non-linear confounds to only keep the most important crossed-term confounds (CROSSED TERMS, 84 confounds). The criteria that we used (described in detail in Section 4.3.2) is that each non-linear confound must pass at least one of two tests, to be included in the confound list:

- 1 Average (across IDPs) of the %VE by the non-linear confound must be higher than the 99.9th percentile of all average % VEs (across confounds).
- 2 Maximum (across IDPs) %VE by the confound must be higher than the maximum of 1 % VE and the 99.999th percentile of all VEs.

4.2.5.4 Date/time-related confounds (120 variables)

Both the acquisition date and time of day could be directly used as confounds, but time or date might represent a confound where the effect on IDPs is highly non-linear and non-monotonic (as a function of date or time), and we are primarily interested in the effect that these variables may have on the IDPs. Therefore, to identify time/date confounds that are rich and flexible but representing average effects in the data, we extracted a set of temporal confounds from the IDP data by sorting the data temporally, smoothing temporally, and performing a Principal Component Analysis (PCA) across smoothed IDPs. This process was performed separately for the acquisition date and for the time of day. We now describe this process in a little more detail.

All previously calculated confounds (487) are first regressed out of each IDP to ensure that we are focusing on time/date effects not already covered by the known above confounds. After regressing out all the previously mentioned confounds, there may still be some variance explained by the acquisition date or its time of day (Karch et al., 2019). This kind of information may be useful for some studies (e.g. drowsiness correlated with time of day), but we would consider it to be a confounding factor in this study.

Next, each IDP is sorted subject-wise according to the corresponding time criteria (either time of day or date), and the IDP values are smoothed using Gaussian-kernel smoothing, with $\sigma = 0.1$, where the units are years for the date smoothing, and work-days (i.e., 13 hours, 7 am - 8 pm) for time smoothing. We then apply PCA on the sorted and smoothed IDPs of the subjects of each site. We retain the number of components that explain 99% of the total variance. The IDP sorting, smoothing, and PCA reduction is applied separately for each scanning site.

This produced 61 components representing the acquisition time, and 59 components representing the acquisition date. The only difference between the generation of the acquisition date and time variables is that the generated acquisition time confounds were also regressed out of the IDPs before generating the date confounds.

4.2.6 Null evaluations (Random Gaussian confounds)

A certain amount of variance in a variable of interest will always be explained by a random variable (i.e., the null scenario). The more variables we have, the more variance would be explained in our variables of interest even by random chance.

Below, (Section 4.3.3), we evaluate how important a confound group is. One way to do this is to calculate how much variance the confound group explains in the IDPs, and also how much unique variance they explain relative to other confound groups⁷. We also check whether our confound groups perform better than the same number of random null variables, as a means to show where our confound groups are useful. These comparisons can be seen in several violin plots.

4.2.7 Correlations between IDPs and non-IDPs variables

In order to assess how unconfounding affects the correlations between IDPs and non-IDPs, we focused on a set of Body variables (224) and Cognitive variables (909)⁸. We extracted this information from the UK Biobank raw data files using FUNPACK⁹. The listing of the variables in each of these two categories as well as the normalization, parsing and cleaning

⁷We calculate the Unique Variance Explained (UVE), in the IDPs, by a group of confounds X, by subtracting from the total variance explained by all confounds (602 variables) the variance explained by all confound groups other than X

⁸For a discussion on the validity and reliability of UK Biobank cognitive tests, see (Fawns-Ritchie and Deary, 2019)

⁹FUNPACK v1.4.1 (MacCarthy, 2019) can be obtained from: git.fmrib.ox.ac.uk/fsl/funpack/

configuration for FUNPACK can be found online¹⁰. We chose these two categories of non-imaging variables as being highly contrasting representative groups: many body variables have strong associations with imaging, and the influence of confounding factors could also be complex and strong; in contrast, associations of IDPs with cognitive variables are much less strong, might easily be problematically dominated by confounding effects, and are likely to be of high interest to researchers using the brain imaging data.

We compared how each group of confounds affects the correlations between IDPs and non-IDP variables (both Body and Cognitive) by finding the P-values of the correlations between each pair of IDP and non-IDP in two settings:

- 1 Regressing out all confounds, other than the confound group in question, from IDPs and non-IDPs.
- 2 Regressing out all confounds from both IDPs and non-IDPs.

We can then compare (Section 4.3.4) the P-values obtained in each setting: if the P-value of the correlation for an IDP - non-IDP pair is higher when a confound group is not used than when it is used, we can be more certain about the importance of that confound group in avoiding spurious correlations.

4.2.8 Additive vs non-additive confound effects

So far, we have only focused on additive confounding effects, but it may be the case that, in addition to linear and non-linear components in a confound, we may have non-additive effects, for example as shown in the final term in the equation:

$$J = I + a * A + b * A^2 + c * A * I \quad (4.1)$$

Where:

I is the true IDP (and our estimation of the true IDP)

J is the measured IDP

A is the confound

a, b, c are scalar factors determining the size of the different confound effects relating to confound A¹¹

¹⁰Code to generate non-IDPs:

[[GIT link](#)]

¹¹So far, we have assumed that confound variable A was estimated perfectly, but as discussed in (Westfall and Yarkoni, 2016), this may not be absolutely true. Nevertheless, their proposed solution, Structural

Up to this point, we have only considered the first part of this equation: $I + a * A + b * A^2$. If all terms are orthogonal, then setting $E = a * A + b * A^2$ and assuming we have correctly estimated E , we can estimate c , that is, we can estimate whether non-additive effects are important for our confounds. If we perform iterative estimation of a , b and c , where c would be calculated from the residual of performing a linear regression, cAI would provide an approximation of how much variance of the IDPs would be explained by the non-additive term. A description of this iterative estimation process can be found in the Appendix, Section A.5.

Results of these analyses on confounds of interest (high % UVE - percent unique variance explained - of IDPs) can be seen in Section 4.3.5. The confounds selected for these analyses were Age, Head Size, and Table Position (4) for subjects in Site 1, because these are, in general, the most important confounds and hence most likely to result in non-negligible interactions with IDPs. We also show results from two motion-related confounds that are examples of smaller amounts of VE and more modest non-additive effects.

4.2.9 Comparisons with other confound configurations (“conventional simple” confounds)

As described above, we have generated a relatively large set of confounds (602 separate variables), which may seem excessive or even impossible to use in small imaging studies. We wanted to compare our proposed set of confounds with a more conventional set of confounds used in typical imaging studies (similar sets of confounds have been used and described in UK Biobank and Enigma Projects ((Miller et al., 2016), (Stein et al., 2012))). Such conventional set of confounds could be:

- 1 Age (1 per site)
- 2 Age squared (1 per site)
- 3 Sex (1 per site)
- 4 Age * Sex (1 per site)
- 5 Head size (1 per site)
- 6 Site (2 variables)
- 7 Head motion (mean relative motion as calculated by FEAT) in resting fMRI (1 per

Equation Models Factor Analysis, requires having multiple measurements of the given confound, but that is not the case in UK Biobank, as several measures of the same metrics were not calculated in general. Therefore, the fact that A is a noisy measure of the true “ A ” means that a , b , & c may be underestimated (i.e., regression dilution)

- site)
- 8 Head motion (mean relative motion as calculated by FEAT) in task fMRI (1 per site)
- 9 Date (number of days when the acquisition happened since the acquisitions started) (1 per site).
- 10 Date squared (1 per site).

We compared this “simple” set of confounds with the same number (29) of random confounds, as described in the previous section. Also, to make a fair comparison with our proposed set of confounds (in the sense that 602 confounds will inherently explain much more variance than 29), we compared the simple set with the first 29 Principal Components of our proposed confound set. Finally, we also compared the simple set of confounds with the PCs explaining 90% and 99% of the variance of the proposed set of confounds (170 and 322 PCs).

For this comparison, we calculated all possible univariate correlations between IDPs and non-IDPs ($3,913 \text{ IDPs} \times 7,247 \text{ non-IDPs} = 28,357,511 \text{ pairs}$) after unconfounding both of them using the different sets of confounds: full set of 602 confounds (ALL), 29 “simple” confounds (SIMPLE), 29 PCs from ALL (PCA-MIN), 170 PCs that explain 90% of the variance of the full set of confounds (PCA-90%), and 322 PCs that explain 99% of the variance of the full set of confounds (PCA-99%). We evaluated how much variance in the IDPs is explained by each set of confounds. We also kept the P-values of those correlations and plotted them in different Manhattan plots similar to (Miller et al., 2016). Finally, with two Bland-Altman plots, we show:

- 1 How the P-values of the correlations between nIDPs and IDPs are reduced when using the full set of confounds (ALL) vs without unconfounding (NONE).
- 2 The difference in the P-values between unconfounding with ALL and the 29 “simple” confounds (SIMPLE).

Results can be seen in Section 4.3.6.

4.3 Results

4.3.1 Non-linear confounds selection

For each of the 3,913 IDPs, we calculated the % UVE by each of the 549 non-linear confounds described in Section 4.2.5.2. Here, the variance is referred to as unique relative to the previous 240 basic confounds (Section 4.2.5.1) that were regressed out of the IDPs (and not meaning unique for all other non-linear confounds). As can be seen in Figure 4.3, the % UVE by non-linear confounds is rather small on average (the largest average % UVE across IDPs being 0.14%), but the maxima can be significantly larger for some combinations of non-linear confound and IDP, e.g. non-linear transformations of age (2.25%), table position (4.44%), or head motion (3.3%). It is also interesting to note that T1w, dMRI and rfMRI Amplitude IDPs were more strongly affected, in general, by non-linear transformations, compared with other classes of IDPs.

An evaluation of appropriate thresholds (described in the next section) for these plots resulted in a subset of 158 non-linear confounds from the original 549.

Non-linear confounds thresholding

We keep non-linear confounds described in section 4.2.5.1 if they comply with either of these 2 rules:

- 1 Mean Variance Explained (across IDPs) of the confound $> 95^{th}$ percentile of the mean VE (0.059)
- 2 Max Variance Explained (across IDPs) of the confound $> \max(0.75, 99.9^{th}$ percentile of all VEs) (1.052)

Here the Variance Explained is “unique” only in the sense of regressing out all the “raw” confounds (section 4.2.5.1) from the IDPs (and not for the other non-linear confounds). We calculated the VE separately by site (i.e. we only used the IDPs of subjects from Site 1 to calculate the VE of a confound for Site 1).

These thresholds were chosen empirically after checking the plots in Section 4.3.1. For $N=5000$ (site 3), once true VE falls much below 1%, estimates of the true VE become (relatively) quite noisy.

If higher-order confounds are included, lower-order versions are automatically included

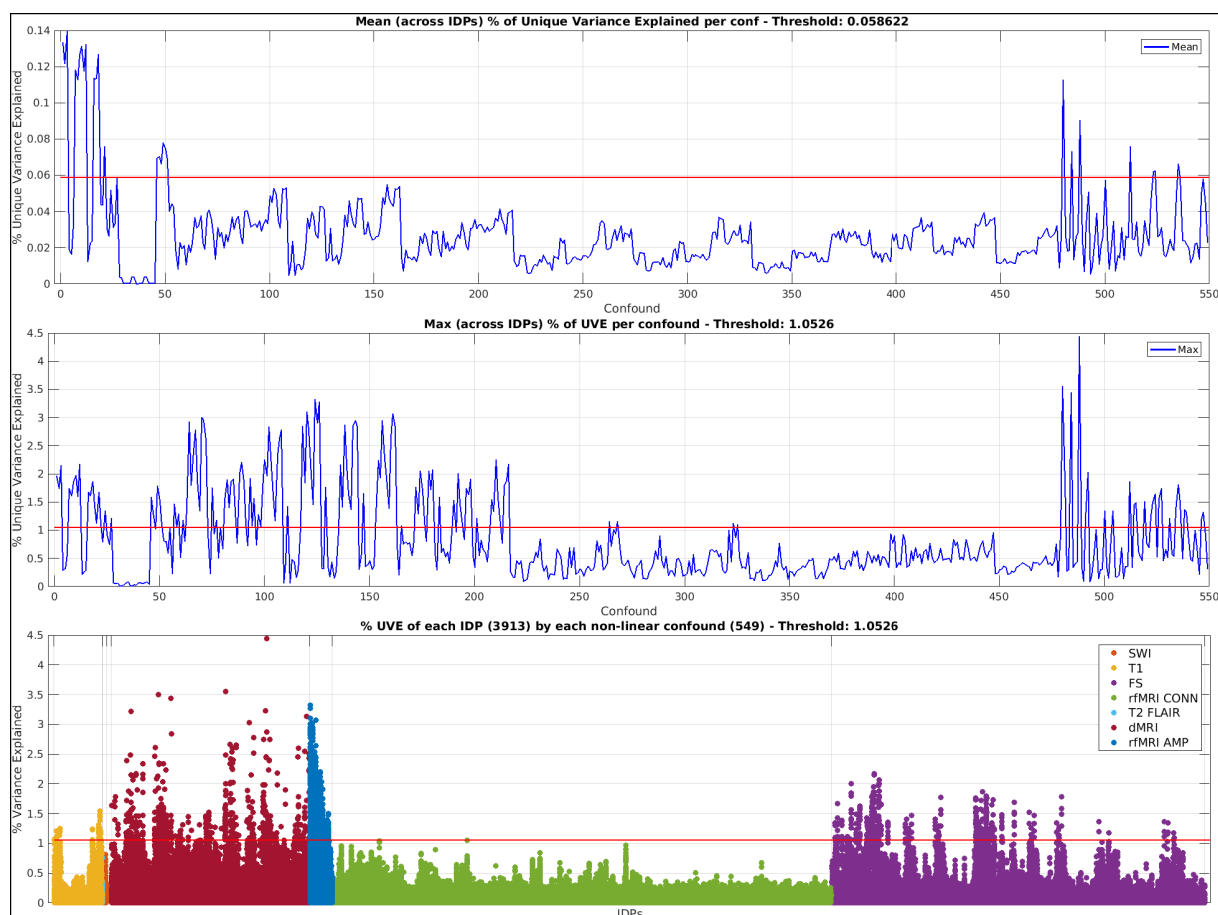


Figure 4.3: *Top* Distribution of the mean (across IDPs) % UVE for each non-linear confound. *Centre* Distribution of the max (across IDPs) % UVE for each non-linear confound. *Bottom* Manhattan plot of the % UVE of each IDP by each non-linear confound, grouped by IDP modality. Calculation of thresholds (red lines in each plot) is described in this section. Interactive versions of these plots, with details of individual results, can be seen at: [\[Top\]](#) [\[Centre\]](#) [\[Bottom\]](#). For Top and Centre plots, the full list of non-linear confounds considered can be seen in [\[LINK\]](#)

even if those did not pass the selection triggers. E.g. if `Age_inormal_squared` is selected, ensure that `Age_inormal` is also selected.

4.3.2 Crossed terms selection

As in the previous section, for each of the 3,913 IDPs we calculated the % UVE (again, unique for the original confounds) by each of the 79,003 crossed-term confounds described in Section 4.2.5.3. Only 26,674 crossed-terms were considered, as the crossing of terms from different sites produces empty confounds. Again, as Figure 4.4 shows, the average (across IDPs) % UVE by crossed-term confounds is small but can be important in some cases, where % UVE goes above 2%. It is also interesting to note that T1w, dMRI, rfMRI Amplitude, and rfMRI Connectivity IDPs have a higher % UVE for crossed-term confounds.

An evaluation of appropriate thresholds for these plots (described in the next section) ended up with a subset of 84 crossed-terms from the original 79,003.

Crossed-terms thresholding We keep crossed-term confounds described in section 4.2.5.3 if they comply with either of these 2 rules:

- Mean Variance Explained (across IDPs) of the confound $> 99.9^{th}$ percentile of the mean VE (0.035)
- Max Variance Explained (across IDPs) of the confound $> \max(1, 99.999^{th}$ percentile of all VEs) (1).

Where the Variance Explained is unique, in the sense of regressing out all the “raw” confounds (Section 4.2.4.1) and non-linear confounds (Section 4.2.4.2) from the IDPs. Again, we calculated the VE separately by site (i.e. we only used the IDPs of subjects from Site 1 to calculate the VE of a confound for Site 1).

These thresholds were chosen empirically after checking the plots in Section 4.3.2. For $N = 5000$ (site 3), once true VE falls much below 1%, estimates of the true VE become (relatively) quite noisy.

If higher-order confounds are included, lower-order versions are automatically included even if those did not pass the selection triggers. E.g. If `Site_x_AcqTime_squared` is selected, ensure that `Site_x_AcqTime` is also selected.

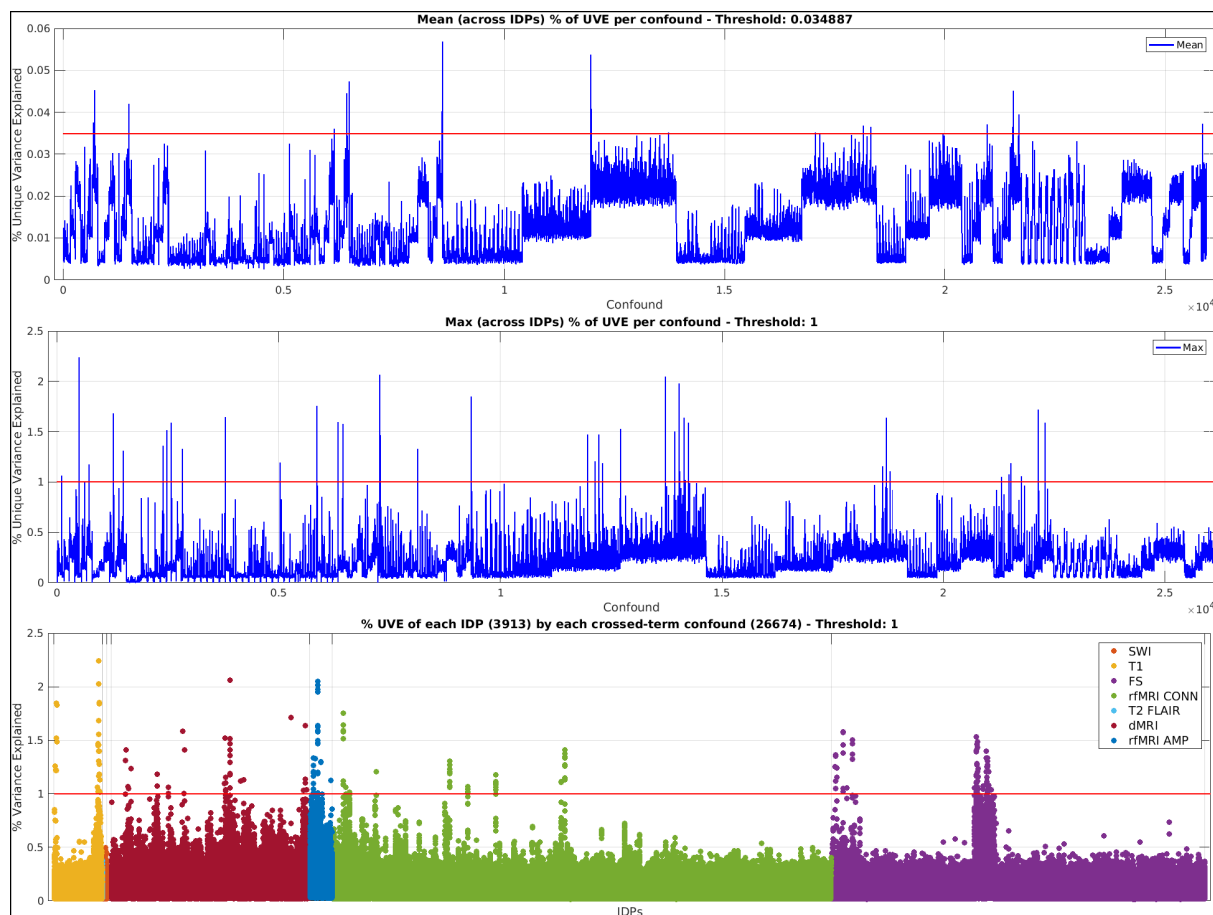


Figure 4.4: *Top* Distribution of the mean (across IDPs) % UVE for each crossed-term confound. *Centre* Distribution of the max (across IDPs) % UVE for each crossed-term confound. *Bottom* Manhattan plot of the % UVE of each IDP by each crossed-term confound, grouped by IDP modality. Calculation of thresholds (red lines in each plot) is described in this section. [\[Top\]](#) [\[Centre\]](#) [\[Bottom\]](#). For Top and Centre plots, the whole list of non-linear confounds considered can be seen in [\[LINK\]](#).

4.3.3 Variance of IDPs explained by confound groups

The percentage of variance explained (%VE) and unique variance explained (%UVE) of the 3,913 IDPs by each group of confounds can be a good indication of the importance of these groups. Figure 4.5 shows UVE for each group (top) or family (i.e., super-group, bottom), and also (in grey) UVE with same-sized sets of random variables. The Y-axis is Log10 of %UVE, and hence “1” means 10% variance of an IDP explained; the (violin-plot) vertical histograms show the distributions across IDPs. VE and UVE for groups/families relate to the variance explained by the relevant set of individual variables considered as a whole together (i.e. are not generated by combining across individual variables’ VE/UVE).

In addition to the expected importance of age, head size and head motion, table position is also notable. Non-linear and crossed-terms are important for some IDPs, as are date and time. Because much of the sex-related confound effects are due to head size, the unique (remaining) effect of sex is smaller than more major confounds like age, head size and head motion, and site effects are on average smaller still. The % UVE of the whole set of 602 confounds is included. For a detailed comparison of the six unconfounding schemes described in Section 4.2.9, see [GLOBAL].

Finally, we also evaluated the % VE and % UVE of between-imaging-site effects in a slightly different way to the above tests (by attempting to adjust for known site effects). Because other confounds are split to give site-specific modelling of each (as described in Section 4.2.9), any between-site mean-effect variations in other factors (including both categorical effects like BATCH and quantitative variables like HEAD MOTION) appear as SITE effects (in addition to any true otherwise unmodelled SITE effects). Hence for this evaluation only, we resorted to the simpler non-site-split versions of all other confounds. This means that between-site mean variations in quantitative variables could now be modelled separately from site, therefore appropriately reducing site effects being estimated here. However, categorical factors that are site-specific or very highly correlated with site (PROTOCOL, BATCH, SCAN RAMP, SCAN MISC, SCAN COLD HEAD, and SCAN HEAD COIL) had to be removed from this evaluation (to avoid these factors soaking up all site variance), which inflates the VE results reported below, compared with the true between-site variation effects not modellable via these known categorical effects.

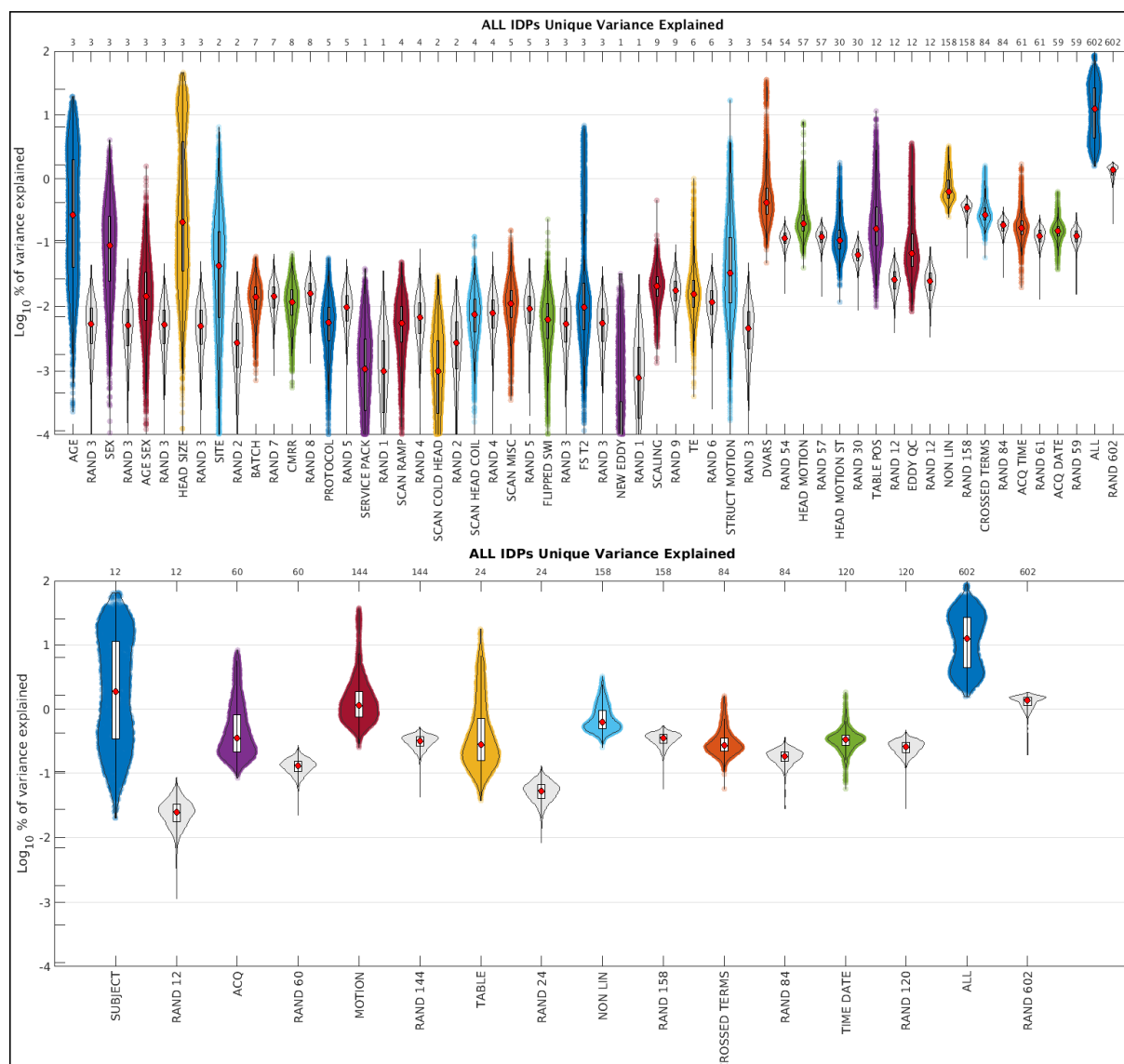


Figure 4.5: *Top* Violin plots with % UVE of IDPs by each group of confounds described in Figure 4.1 [UVE Top]. For a similar figure showing the VE instead of the UVE: [VE Top]. *Bottom* Violin plots with the % UVE of the IDPs by each family of confounds described in 4.1 [UVE Bottom]. For a similar figure showing the VE instead of the UVE: [VE Bottom]. Appendix (Section A.7) shows the same data detailing the variables by IDP modality. Light grey violin plots show the % VE or % UVE explained by the same number of random variables (each set of matched-size random null variables is generated uniquely, hence the small variations between same-sized RAND groups). An interactive version of all these violin plots where the reader can verify the exact VE and UVE of each IDP explained by each confound group or family, in total or by IDP modality, is available at [LINK].

The results show that the upper limit on site variability (%UVE) has a 75th percentile close to UVE=1%, which is smaller than any of the confound families shown in Fig 4. UVE reaches 10% for the most strongly affected IDPs, but the majority of the IDPs are considerably lower. We conclude that it is worth unconfounding for site, but that site effects are minimal for most IDPs. That VE is larger than UVE reflects the fact that some (modelled) confounds are correlated with site; for example, table position has now been optimised and is set consistently across all sites, but early scans for site 1 had a wider range of table positions. We would therefore not consider this a core “site effect”, given that it is controllable and modellable.

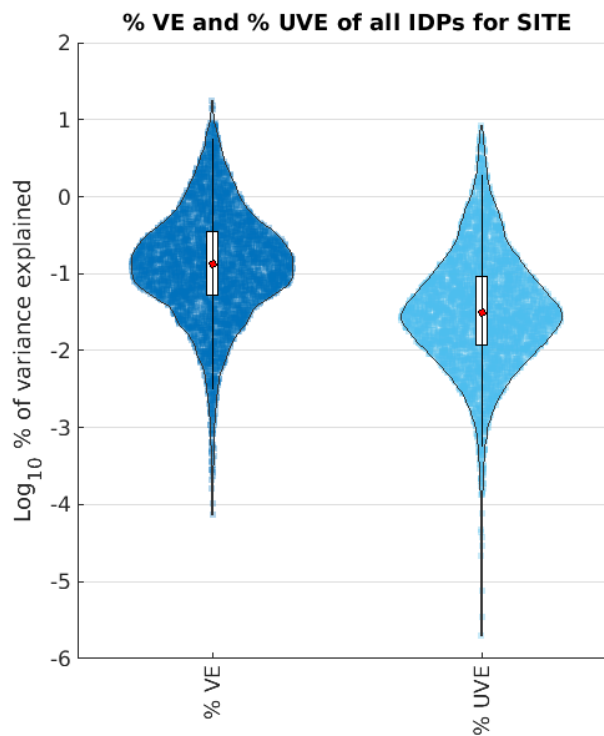


Figure 4.6: % VE and % UVE of IDPs by Site confound (distributions are across IDPs). The individual mini-blobs (just visible in blue here, and more obvious in sparser plots below) reflect individual data points of the distributions. [\[VE_UVE_SITE\]](#)

4.3.4 Effect of unconfounding on IDP-nIDP correlations

For each confound group, we compared how the P-values of the correlations between IDPs and Cognitive and Body variables change when unconfounding both IDPs and non-IDPs with all confounds, versus unconfounding them with all confounds except for the confound group of interest. By doing this, we can see how each confound group “uniquely” affects the correlations.

All Bland-Altman (BA) plots are shown in the Appendix (Section [A.8](#)). In Figure [4.7](#)

we include a few exemplar BA plots. AGE affects almost all correlations, as expected; the same is true of HEADSIZE. Unconfounding TABLEPOS does not strongly affect the correlations between Cognitive and IDPs but has a stronger effect on correlations between Body variables and IDPs. As mentioned above, unconfounding CROSSED TERMS does not strongly affect the correlations, although there is a small systematic (overall trend) effect. There is almost no effect of SERVICE PACK, as one would hope, as this refers to what should be minor Siemens software upgrades. Finally, we can see an example of Berkson's Paradox (discussed further below) on how unconfounding DVARS confounds increases the significance of the correlations between Body variables and IDPs.

4.3.5 Importance of non-additive terms

We evaluated the importance of non-additive terms as defined according to equation (1). The results in Figure 4.8 show how the non-additive component of the eight studied confounds have low importance to explain the variance of some IDPs in terms of non-additive confound effects (all distributions are across IDPs). For each confound, we can see the correlation between the measured IDP and the estimated IDP (left), % VE for the linear, quadratic and non-additive terms. We also compare with the % VE of a random variable (null). The non-additive term is generally smaller than the quadratic term, but not by a large amount, and it is generally larger than the null scenario (variance explained by a random regressor).

4.3.6 Effects of unconfounding with different sets of confounds

We found that different sets of confounds (except for PCA-MIN) explained more variance of IDPs overall than the SIMPLE set of confounds (Figure 4.9). Here, ALL had 11.14 mean %VE, SIMPLE had 4.38 mean %VE, PCA MIN had 3.22 mean %VE, PCA 90% had 7.44 mean %VE, and PCA 99% had 9.5 mean %VE. We also show the general reduction in P-values of the correlations for IDPs vs non-IDPs when unconfounding with the full set of confounds (ALL) and the reduced set of confounds (SIMPLE) with two Manhattan plots (Figure 4.10). Similar plots can be seen in Section A.10 of the Appendix and we can see the different number of associations passing Bonferroni correction for each set of confounds:

- No unconfounding: 1,665,274
- PCA MIN: 984,142

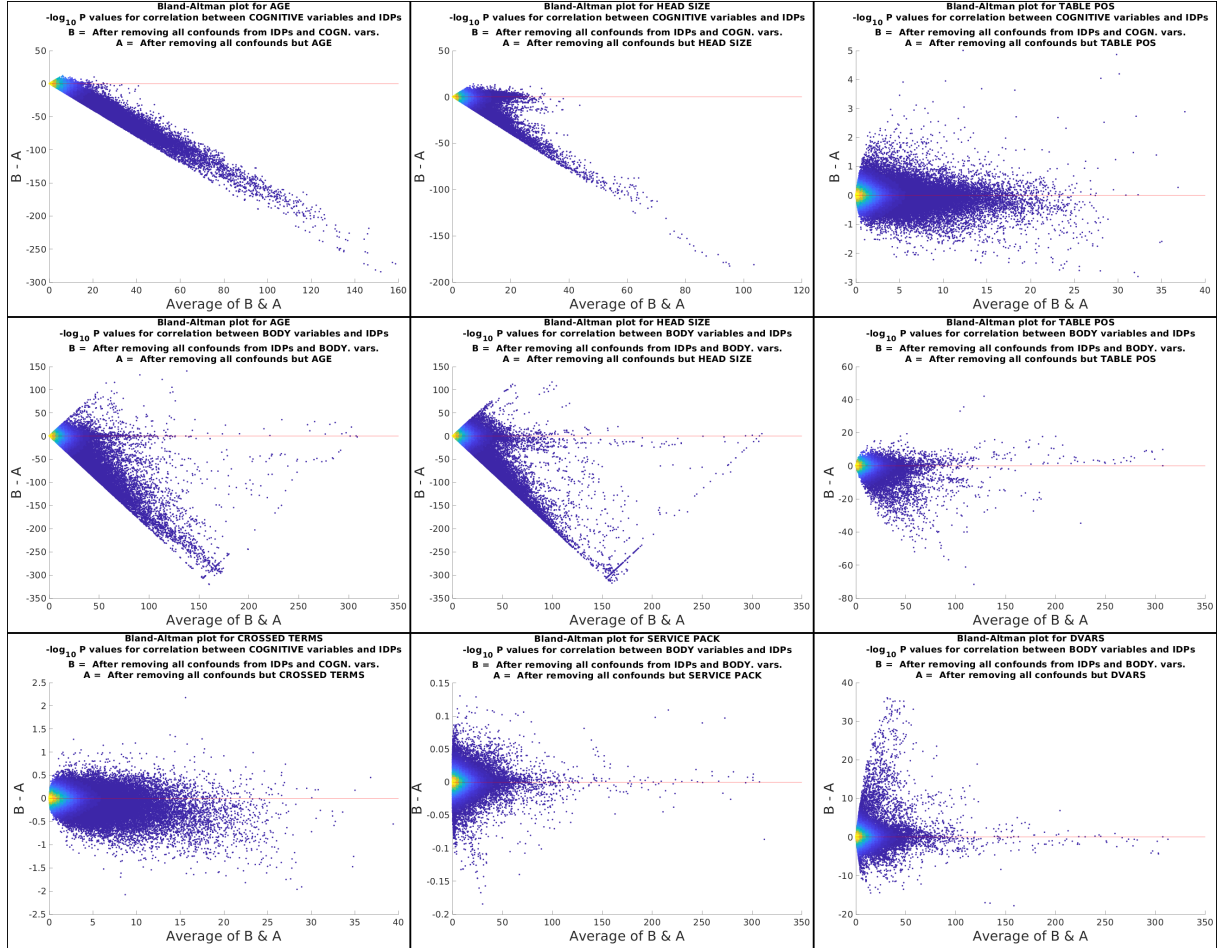


Figure 4.7: We show here a subset of all the Bland-Altman (BA) plots produced, which illustrate how correlations of IDPs with Body and Cognitive variables are affected very differently by the unconfounding. In these plots, a situation where a confound group does not strongly affect the correlations would appear as a horizontal cloud of points around $y=0$ (meaning no substantial difference between A and B). Where the cloud of points leans heavily towards negative y , this means that using that confound group reduces the significance of correlations (implying that the correlations were spurious). If the cloud of points leans heavily towards positive y , this implies a case of Berkson's Paradox, particularly where values in A are close to zero. The remaining BA plots can be found in the Appendix (Section A.8). Interactive versions of all BA plots, where the reader can verify the exact change in P-values and the IDP / non-IDP pair that each point represents can be found in [\[LINK\]](#).

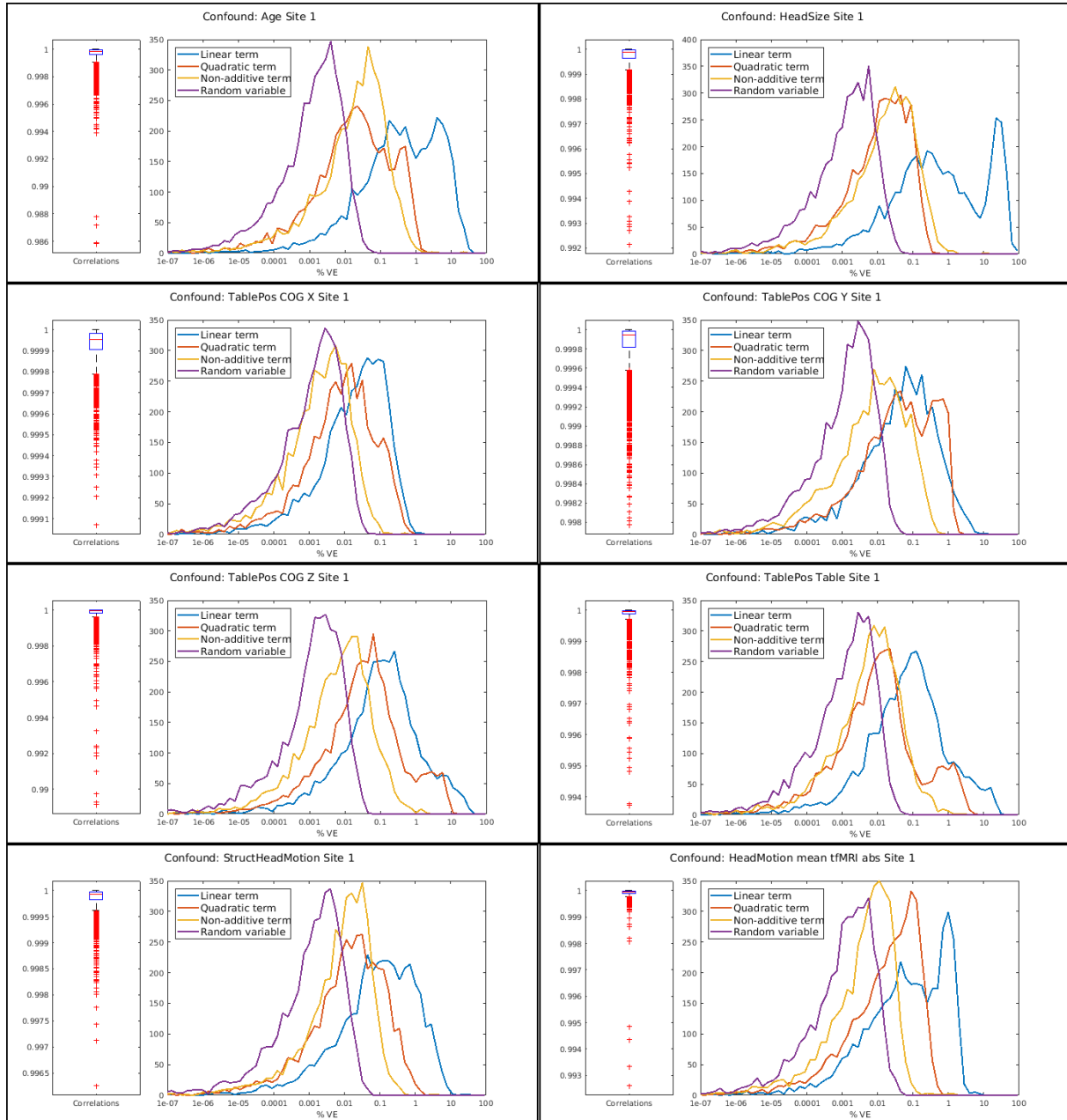


Figure 4.8: Effect of modelling non-additive terms. Each panel shows a different confound: (Left) Correlation for the measured IDP (J in equation (1)) with the estimation of the true IDP (I in equation (1)). The boxplot distributions are across IDPs. (Right) Histograms (distributions across IDPs) of the % Variance explained for IDPs by the Linear term, the quadratic term, the non-additive term and a random variable for null comparison.

- PCA 90%: 140,963
- SIMPLE: 105,122
- PCA 99%: 58,949
- ALL: 53,995

Finally, we compared those two unconfounding settings directly via a Bland-Altman plot (Figure 4.11a); this suggests that for many cases, more “complete” unconfounding improves sensitivity for finding significant associations.

4.3.7 Date and time confounds

The complete set of generated Acquisition Date and Acquisition Time confounds can be found in A.6 of the Appendix. Interpretation of the meaning of these data-driven confounds may be complicated but in some cases, possible. One clear example is the first temporal component (See Figure 4.12.A). This component was highly correlated (Figure 4.12.B) with one of the resting fMRI Node Amplitude IDPs (Figure 4.12.D) which are known to correlate with drowsiness ((Stoffers et al., 2015) and (Bijsterbosch et al., 2017)). The periodicity (6 events per day) can be analysed in light of the number of subjects per day that are scanned in each site (roughly, 18 per day, being Site 1 the most consistent). This periodicity may be related to how much the subjects have rested and how this is only noticeable after the heavy smoothing applied before PCA (Figure 4.12.D: The periodicity is hard to see in pre-smoothed IDP data).

According to the UKB imaging centre workflow, participants circulate around the imaging centre in triplets. In each triplet, the three subjects perform different activities in different orders (e.g. subject 1 is being scanned in the brain MRI while subject 2 is performing physical exercise and subject 3 is being scanned in the cardiac/abdomen MRI). Therefore, one subject out of three will have their brains scanned immediately after doing exercise, while another will be scanned after “resting” in the cardiac scanner; this may affect the results.

4.4 Discussion

4.4.1 General considerations

In this study, we have confirmed the importance of known confounds such as age, sex, head size and head motion, while also showing that in the UK Biobank imaging data, the

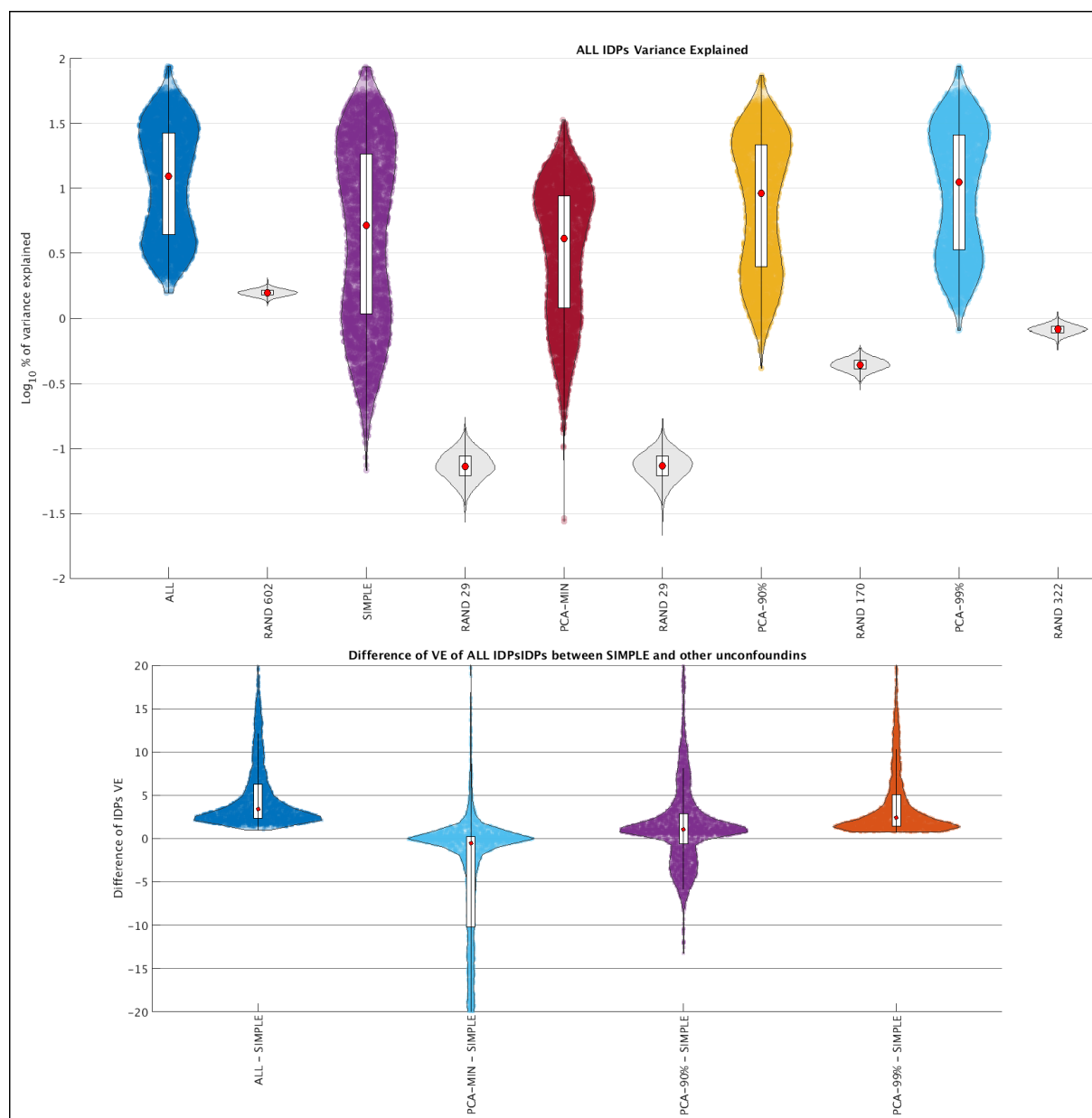


Figure 4.9: *Top* Violin plot with the amount of variance of all IDPs explained by different sets of confounds: ALL (the full set of 602 confounds that we have developed in this work), SIMPLE (a more common set of confounds used in most studied and described in Section 4.2.9), PCA-MIN, PCA-90% and PCA-99%: Three sets of Principal Components described in Section 4.2.9) obtained from ALL. The first has as many components as confounds in SIMPLE (29), the second has the number of components that explain 90% of the variance of ALL (170), and the third has the number of components that explain 99% of the variance of ALL (322). Each of these sets of confounds is compared with a set of random confounds of the same size. An interactive plot (where the reader can check how much variance is explained by each confound in each set) can be seen in [\[GLOBAL_ALL\]](#). *Bottom* Violin plots showing the distributions of paired-differences in VE of all IDPs, comparing the SIMPLE set of confounds and the other sets of confounds.

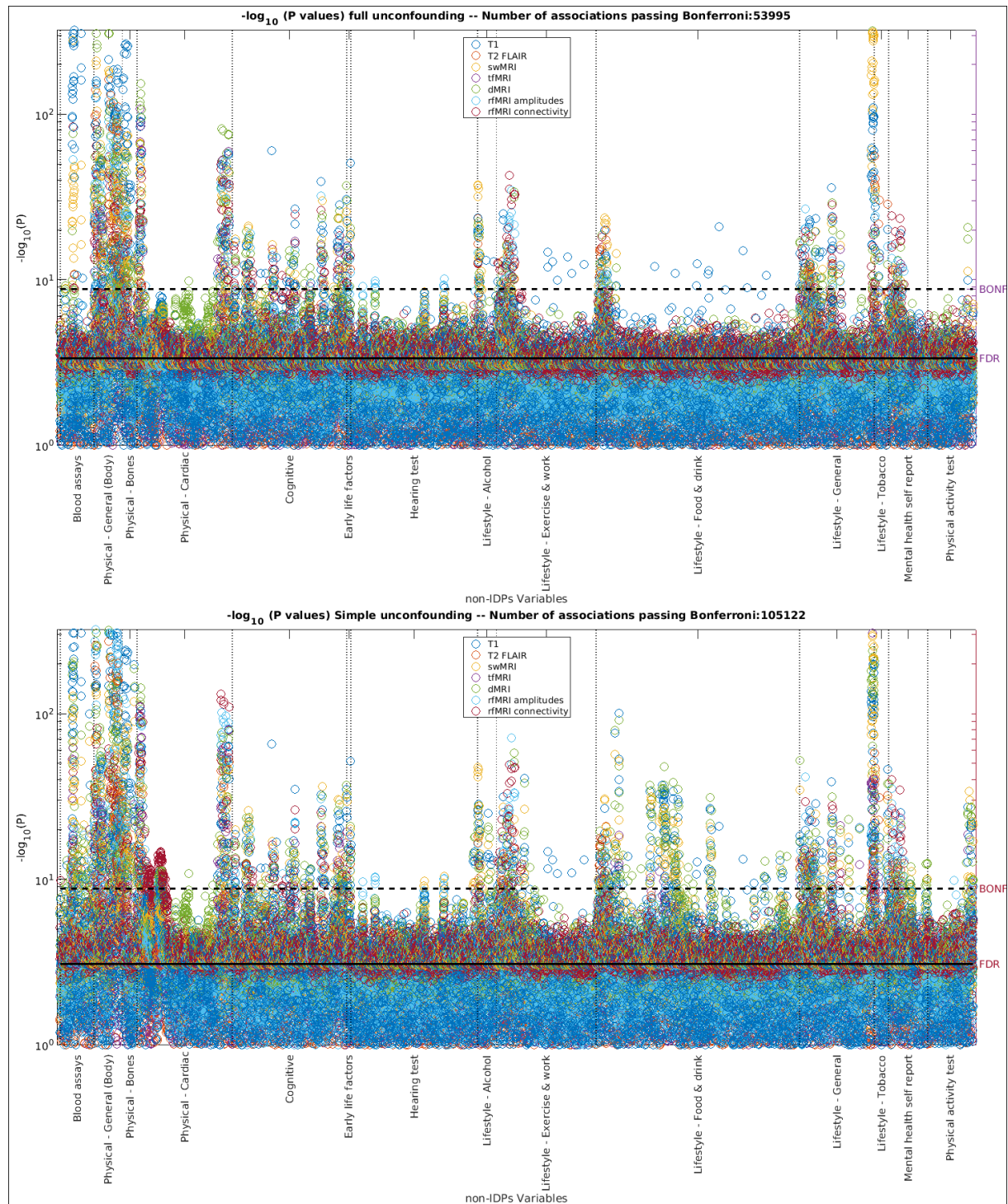


Figure 4.10: *Top* Manhattan plot showing how the correlations between IDPs and non-IDPs are affected by unconfounding with the whole set of confounds [Top]. *Bottom* Manhattan plot showing results after unconfounding with the SIMPLE set of confounds [Bottom]. The main difference between the plots is that the number of correlation tests between IDPs and nIDPs passing Bonferroni correction is greatly reduced using the full (ALL) unconfounding (53,995) than when using SIMPLE unconfounding (105,122). This would imply that half of the significant (Bonferroni-passing) correlations using SIMPLE unconfounding may not be meaningfully significant. Similarly: [No unconfounding] [PCA-MIN] [PCA-90%] [PCA-99%].

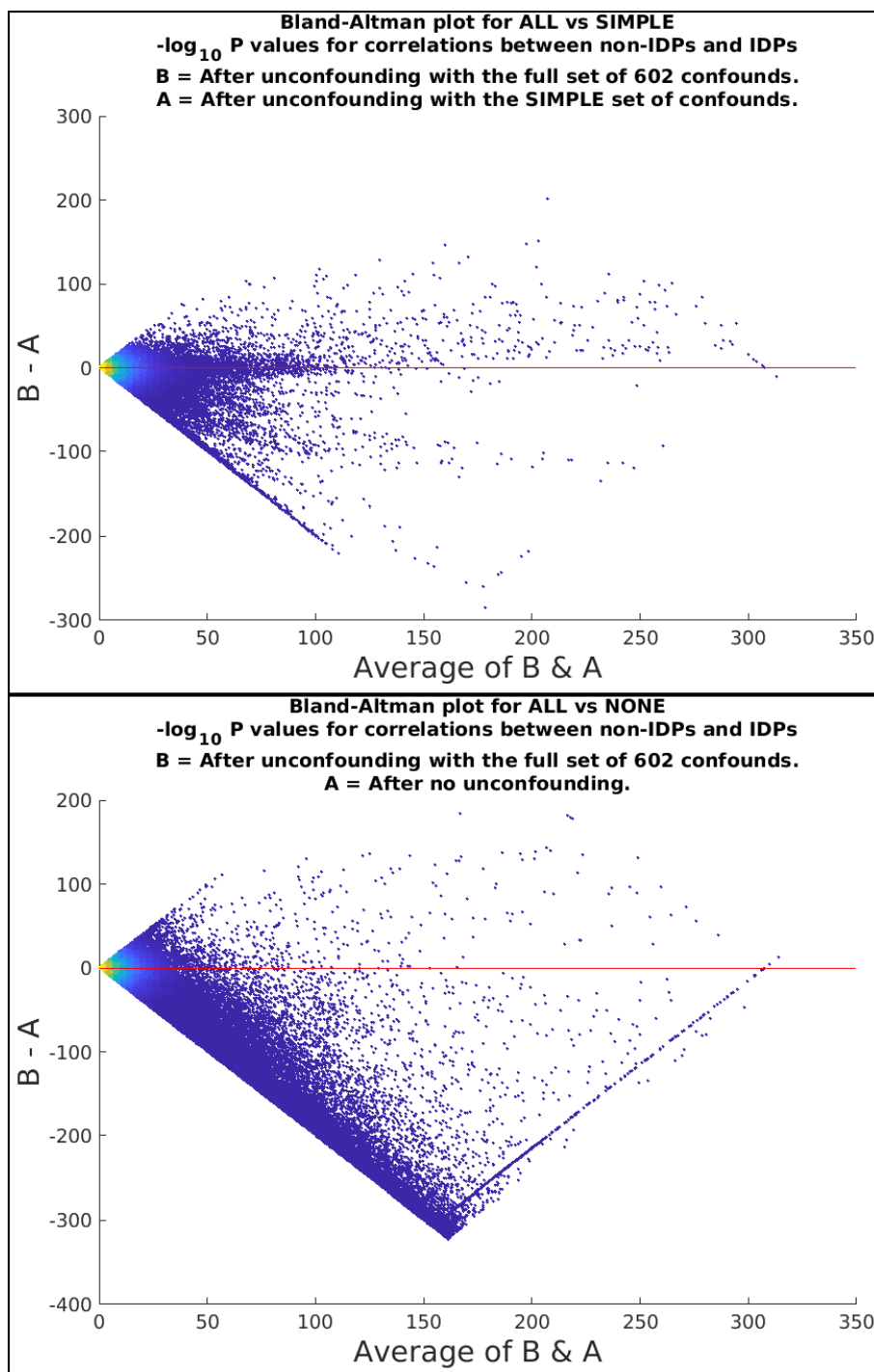


Figure 4.11: *Top* BA plot to show the difference in P-values for the correlations between IDPs (3,913) and non-IDPs (7,247) when using two different unconfounding settings: full set of 602 confounds (ALL) and “common” set of 25 confounds (SIMPLE). *Bottom* BA plot to show the difference in P-values for the correlations for IDPs and non-IDPs when unconfounding with the full set of 602 confounds (ALL) and without any unconfounding. The diagonal line (bottom-right) is due to some correlations without any unconfounding (A) having a smaller P-value than the numerical precision limit. Note that adding more confounds might make P-values go in either direction: it might increase sensitivity to real effects (which is likely what we see in A), or it might decrease the strength of correlations because fake associations (caused by the confounds in the data) go away (B).

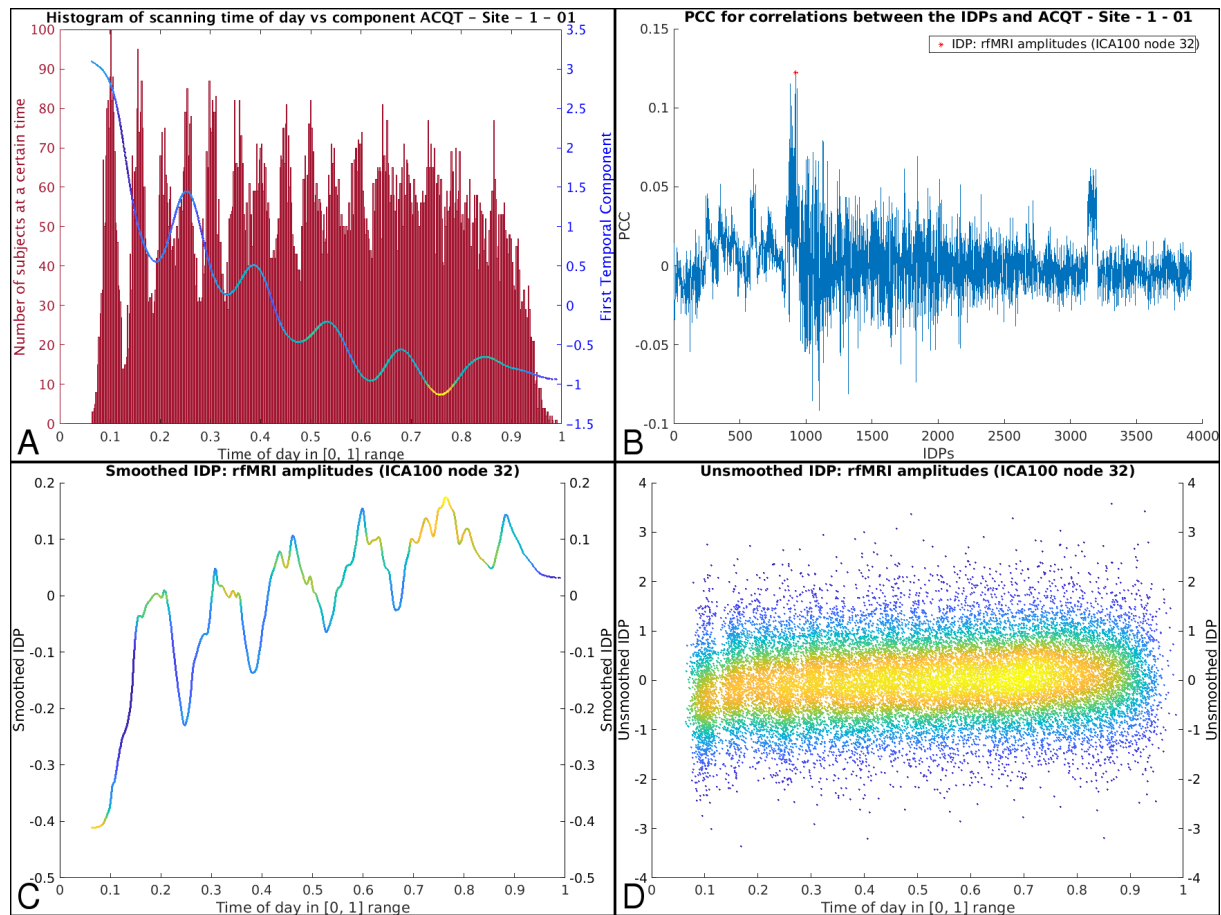


Figure 4.12: *A* First Principal Component (PC) for the Acquisition Time confounds for Site 1, along with the histogram (in red) of the acquisition times of all Site 1 subjects, where the main peaks (of “dominant” imaging start times) can be easily identified. The PCA component is the strongest time-drift effect (across all IDPs) that is not already removed by other known confounds. *B* Plot of all the correlations between this first PC and each IDP. The two most strongly correlated sets of IDPs are rfMRI node amplitudes, and T1 intensity contrast across the white-grey cortical boundary; IDP rfMRI Amplitude (ICA 100 node 32) is the most correlated. *C* Smoothed (moving average with a span of 1000) of just this IDP over time, which is clearly tending towards the first PC. *D* The same IDP, without temporal smoothing (one point per subject).

position of the scanner table, as well as many other acquisition parameters and configurations, have a non-trivial influence on the data. Although UK Biobank brain imaging data is of high overall quality and homogeneity, the extremely high subject numbers means that even small confounding effects can cause statistical problems (in particular, raising the danger of finding false-positive associations).

In developing this set of confounds, we considered including different quality metrics, such as the “Entropy” of the T1; however, we chose to omit these here because they could be closely entangled with related IDPs such as WMH volume (and other valid “signal” effects). Similarly, we initially included a binary variable called “Repeated T1”, indicating whether the acquisition of the T1 image was repeated more than once. We initially speculated that this could be used as a proxy for structural head motion, but we instead decided to represent structural motion more thoroughly in the `STRUCTMOTION` variable, explained in Section 3.3.3 of the Appendix.

Some considerations may need to be made regarding the basic confounds that we studied. For example, we considered whether it would be more appropriate to use brain size instead of head size as a possible confound. This is a complicated question that may deserve exploration in the future. It may also be interesting to include FreeSurfer-head/brain size scaling in this analysis, as these are calculated in a different way than we have done here.

It is also important to note that most IDPs have not been normalised for head size before unconfounding. In general, it is only sensible to scale IDPs by a head size scaling factor when the IDP in question is a raw volumetric measure. For other types of IDPs, normalising (scaling) by non-demeaned head size is likely to induce head size confounds in the data (this is similar to the general danger of inducing confound effects when regressing out confounds that have not been demeaned). In general, the safer way to approach unconfounding is to regress out demeaned confound regressors.

The generation and processing of non-linear confounds and crossed terms shown in this work is not necessarily intended to be the exact rigid recommendation for future studies. Some of the % VE thresholds described in Section 4.3, though driven by inspecting data distributions, are somewhat arbitrary. It may be that cross-validation (or other methods for model-order selection) should be used to obtain a more objective selection of confounds to keep. However, the presence of complicated patterns of shared variance between differ-

ent confounds would complicate the efficient application of such methods, given the large number of potential confounds. Our main intention here was to show how these terms may be important and illustrate their effect on downstream analyses. In the same way, we have shown that non-additive terms may also play a non-trivial role, but finding the optimal way to identify and deal with these will depend on the research question being asked, and the level of unconfounding conservativeness desired.

Finally, some of the results presented here may have been affected by the fact that we have only included subjects that passed the QC process described in (Alfaro-Almagro et al., 2018) (i.e., subjects with a T1 deemed “unusable” were already excluded). Confounds such as STRUCT MOTION and HEAD MOTION may have been found to be significantly more important to the unconfounding process, had those subjects not been left out.

4.4.2 Confounds and non-IDP variables

The rich relationships (in terms of VE and UVE) of the confounds with non-IDPs are beyond the scope of this study, although we have illustrated in many Bland-Altman plots how the confounds affect associations between IDPs and non-IDPs. For example, the importance of head motion as a confound cannot be overstated for most studies, but this confound may also be correlated with important effects of interest. Correlations between motion-related confounds and health-related variables are to be expected, even in healthy subjects - for example, (Bijsterbosch et al., 2017) identified a relationship between head motion and sleep quality. Our results show that DVARS metrics are correlated with ECG load during exercise bike activity, and also with resting-state fMRI node amplitude. In fact, all resting-state node amplitudes correlate somewhat similarly with DVARS, but sensory-motor nodes correlate more strongly than cognitive nodes (consistent with (Bijsterbosch et al., 2017)). Hence, there is possibly an underlying physiological factor which is manifesting itself in health-related variables, and in resting-state connectivity and fMRI motion. Therefore, while in this study we are using DVARS metrics as confounds, they could potentially find use as health biomarkers.

Some non-IDPs, such as blood pressure, bone density, height and weight, can be strongly associated with IDPs ((Miller et al., 2016), (Smith and Nichols, 2018)), and might be considered for use as confound variables. These are likely to correlate with confounding processes in the imaging (for example, weight correlates with head motion), but are

also likely to reflect physiological processes of interest partially. Subject age is a complicated example of this, mediating large amounts of between-subject variance that could be considered signal, noise or both. Here we focused primarily on confounds that we could derive from the imaging data (although some of our imaging-derived confounds are obvious proxies for non-imaging measures, e.g. head size is a proxy for height).

4.4.3 Number of confounds

The motivation of this chapter is to explore approaches to controlling for confounds in large scale prospective brain imaging studies; we have aimed to provide recommendations for the types of confounds that should be used in such studies. For simple focussed correlation studies, researchers may want to use simpler or more specific versions of the confounds we have proposed; this raises the question of how we can reduce the number of confounds a researcher should use.

As was shown in Section 4.3.6, using a small number of confounds may suffice, especially if those confounds are specific enough. For example, the SIMPLE confound set explains more variance in IDPs than the PCA-MIN confound set. This is due to the SIMPLE set being, in general, (almost by definition) focussing on the strongest (most important) compact set of confounds; ALL expands on this with a much larger number of (in many cases more subtle) confounds, and PCA-MIN is an unsupervised reduction of ALL, and hence is not expected to be as focussed at accounting for the strongest confound effects as SIMPLE. Nevertheless, on the whole, the SIMPLE confounds do not explain (on average) as much variance as the ALL, PCA-90%, or PCA-99% sets (Figure 4.9). These comparisons may be different for different modalities, as can be seen in Section A.7. For example, the most important confounds for T1, T2 and FS are head size, age and sex. Those are much more important than anything else, and they are included in both ALL and SIMPLE. That is why for those modalities, the %VE are not so different.

Although the investigation of the effects of unconfounding on associations between IDPs and nIDPs can be informative (for example, this can help eliminate confounds that make no difference), it can be hard to judge whether changes in IDP-nIDP association P-values are beneficial or detrimental, as there is, in general, no ground truth available. For example, “good” unconfounding might reduce P (increase significance) by correctly removing noise that is not shared between an IDP and an nIDP, or could increase P by

correctly removing noise that is shared. On the other hand, “bad” unconfounding might increase P by over-aggressively removing good signal, or might incorrectly decrease P due to Berkson’s paradox (i.e., regressing out an inappropriate confound - see below).

4.4.4 Berkson’s paradox

One problem that may arise when applying unconfounding is known as the Berkson’s paradox ((Pearl et al., 2009), (Zhang, 2008)). This effect occurs when we adjust two independent variables for a potential confound that was actually a consequence of the independent variables. In this case, a spurious association between the independent variables can be incorrectly induced.

Even without this known mathematical problem, such unconfounding does not make sense to apply, if the “confound” factor was actually caused by (and not a causal factor feeding into) the two variables being tested for an association. Examples of mostly safe confounds are age, sex, and genetics, as these are “causal” factors unlikely to be “caused by” other imaging and non-imaging variables of interest.

A possible illustrative example of Berkson’s paradox in our results could be when considering the association between the fMRI task activation (which involves visual cortex stimulation and button pushing) and the body variable of waist size. With unconfounding that excluded head motion confounds, the association was very weak ($-\log_{10}(P) = 0.55$), while after also unconfounding for head motion, the association was stronger ($-\log_{10}(P) = 2.84$). It seems quite plausible that both variables have some “input into” head motion, and hence the unconfounding increases the IDP-nIDP association from being close to zero to being significant.

4.4.5 Sex-split unconfounding

As mentioned in Figure 4.1 and detailed in Section 4.2.5.1, all confounds were modelled to be orthogonal to site (i.e, a site-specific version of each confound was created). There may be other cases where confounds need to be split; for example, if a confound affects males and females differently (with different strengths of confounding effect), sex-split confounds might be necessary.

There may be additional reasons why such splitting is necessary. For example, in cases where the raw data (e.g. some disease outcomes) is all-zeroes for one sex (either because

one sex has all zeros “by definition” - like for testicle-related disease in females, or because the subset of subjects being processed happens to be all zeroes for one sex), then statistical problems can arise. If a confound is not sex-separated, the all-zeroed-values (i.e., for one sex) will likely get the confound induced in them (i.e. are no longer all zeros). Note those induced values are, of course, a perfect copy of the confound. This becomes a problem if the study then tries to correlate the unconfounded variable with something else which has confounds in it (or if sex-split associations are tested for). The use of split confounds avoids the problem.

4.4.6 Availability of the confounds

We define “source” confound variables as those fundamental variables (e.g., head size, age, sex, imaging site) from which the entire set of confounds described in this work can be derived; ideally, the source variables will all be UKB variables. They might have been derived from image-level processing (i.e., requiring the NIFTI image-level downloads), the DICOM headers, or other information from the imaging centres such as scanner hardware service/change dates. We define “non-source” confound variables to be those that are easily derived from the source variables (for example, non-linear versions of source variables, and interaction terms). Our aim is for it to be easy for researchers to obtain and work with the source variables, and then use our code from this work (or their own alternative code) to derive non-source variables.

The most important “source” confound variables (including variables needed to form the “SIMPLE” category) are already directly available from UK Biobank (and we list these, along with their UKB variable ID codings in Table 4.1). The remaining source variables that have been identified in this work will be sent to UKB to be made available for researchers shortly. These variables are also listed in Table 4.1 (with UKB variable codings denoted with “TBD”, to be decided).

Source variables are therefore not already processed in all of the more complex ways described in this chapter (e.g., split to create site-specific variables, quantile normalised, used to create higher-order/interaction effects, with categorical confounds turned into multiple binary indicator variables, and with date/time drift variables estimated). This is in part because such derived, non-source, confounds depend on the specific set of subjects that are used in a particular unconfounding process. To carry out site-specific (etc.) con-

found modelling will require the researcher to follow processing similar to that described here. We here provide freely available code to both generate the source variables from all the different sources¹² and to process them to generate the confound sets used in this chapter¹³

4.5 Conclusions

In this chapter, we have developed a set of brain imaging confounds and have tested their importance in the context of their effect on IDPs and on association studies between imaging and non-imaging variables, through investigation of the Variance Explained and the Unique Variance Explained, as well as how each group of confounds affects the data. We have generated a large number of interactive plots that can be explored in order to get a fine-grained idea about how each group of confounds are related to each other, affect the IDPs, or affect the relationship between IDPs and non-IDPs. We have shown how to study the importance of a possible confound, what confounds may be useful, and possible ways of reducing their number. We have identified a set of confounds that could be useful in the context of UK Biobank brain imaging data (which covers six structural and functional modalities, and already has 40,000 participants' data released). A centralised description of these imaging confounds (including code for generating the confounds from the subject-level data disseminated by UKB, and code for applying the confounds) is available at www.fmrib.ox.ac.uk/ukbiobank/confounds/ and this will be updated as more subjects' data are released (and further confounds are identified) in an ongoing way.

¹²https://git.fmrib.ox.ac.uk/falmagro/ukb_unconfound/-/tree/master/generate_initial_data

¹³https://git.fmrib.ox.ac.uk/falmagro/ukb_unconfound/-/tree/master/confound_processing

Chapter 5

Discussion and future work

5.1 Summary

In this thesis, we have shown our research to produce reliable and well-curated brain imaging information in one of the largest and richest publicly available biological datasets. We have shown our efforts to process the data, guarantee its quality, and give recommendations on how to use it better.

- In Chapter 1, we have introduced UK Biobank, describing all its components with special emphasis on the brain imaging part, the modalities that form it, and the changes in the acquisition protocol that it has undergone.
- Chapter 2 describes the development of our Processing pipeline, the resulting processed images, the developed IDPs, and the investigations to assess the optimal steps in this pipeline as well as changes in the acquisition protocol.
- Chapter 3 describes the QC (Quality Control) methods we studied to make sure that the publicly distributed data can be used by researchers. We specifically focus on methods to assess the quality of T1w images and their non-linear registration to the atlas. We also describe how we are performing QC on FreeSurfer outputs.
- In Chapter 4, we have shown how we have approached the problem of dealing with confounds in large cohort studies. We proposed a set of possible confounds and explained a methodology to evaluate them, not just in UK Biobank, but in other brain imaging studies. We showed how to test their importance in the context of their effect on IDPs and on nIDPs, and possible ways of reducing their numbers. We have also shown how this set of confounds may be more useful in some studies than the normally used, simpler, set of confounds, and why this is important in large cohort studies. Effectively, we are showing the best way to use the data we have produced in the previous steps.

We hope this work can be useful for the whole neuroimaging community.

5.2 Impact on neuroscience research

The work here described has facilitated many other researchers to develop their own studies by using either the processed images we produced, the IDPs, or the confounds. This is a brief list of some of the publications that use this data:

There is a variety of studies looking at **differences in brain volumes** (total or regional) between individuals, mainly for health research purposes, but also to get a better understanding of the brain. Most of these studies use the different volume IDPs we have generated (e.g. [Ambikairajah et al. \(2020\)](#), [Dekkers et al. \(2019\)](#), [Fjell et al. \(2020\)](#), [Gale et al. \(2020\)](#), [García-García et al. \(2019\)](#), [Garfield et al. \(2020a\)](#), [Garfield et al. \(2020b\)](#), [Gheorghe et al. \(2019\)](#), [Gray et al. \(2020\)](#), [Hamer and Batty \(2019\)](#), [Kiesow et al. \(2020\)](#), [Lee et al. \(2019\)](#), [Nobis et al. \(2019\)](#), [Norbury \(2019\)](#), [Rudner et al. \(2019\)](#), [Salminen et al. \(2018\)](#), [Schulz et al. \(2020\)](#), [Shen et al. \(2017\)](#), [Taebi et al. \(2020\)](#), or [Warland et al. \(2020\)](#)). Some studies use the NIFTI images we generated and run their own processing to generate the volumes to study (e.g. [Alnæs et al. \(2019\)](#), [de Lange et al. \(2019b\)](#), [de Lange et al. \(2020b\)](#), [Eshaghi et al. \(2019\)](#), [Kong et al. \(2019\)](#), [Le Guen et al. \(2019a\)](#), [Majrashi et al. \(2020a\)](#), [Majrashi et al. \(2020b\)](#), [Van der Meer et al. \(2018\)](#), [Nave et al. \(2019\)](#), [Norbury \(2020\)](#), [Stolicyn et al. \(2020\)](#), [Vinke et al. \(2019\)](#), [Zhao et al. \(2019a\)](#)). Finally, a group of studies look at differences in brain volumes, but combine this data with a wide range of our IDPs (e.g. [Cox et al. \(2019a\)](#), [Daviet et al. \(2020\)](#), [Ritchie et al. \(2018\)](#), and [Wiberg et al. \(2019\)](#)).

Some studies take a look at brain **white matter**, either related to white matter hyperintensities from T2 FLAIR images or structural features from the white matter from diffusion data. To do this, some studies use a wide range of IDPs (e.g. [Cox et al. \(2016\)](#), [Croall et al. \(2020\)](#), [Liu et al. \(2018\)](#), [Lyall et al. \(2019\)](#), [Mollink et al. \(2019\)](#), [Park et al. \(2020\)](#), and [Rutten-Jacobs et al. \(2018\)](#)). Other studies are more focused and just use either dMRI IDPs (e.g. [Cox et al. \(2019b\)](#), [Dekkers et al. \(2019\)](#), [Shen et al. \(2019\)](#), or [Suzuki et al. \(2017\)](#)), or brain volume IDPs (e.g. [Shen et al. \(2017\)](#), [Traylor et al. \(2019\)](#), or [Tsang et al. \(2019\)](#)). Finally, some studies run their own processing using the images we generated (e.g. [Beaudet et al. \(2020\)](#), [Sundaresan et al. \(2019a\)](#), [Veldsman et al. \(2020\)](#), [Zhao et al. \(2020\)](#), [Zhao et al. \(2019b\)](#)).

Brain connectivity (either structural or functional connectivity) is a growing field that

has also benefitted from UK Biobank data. To do this, some studies use our rfMRI IDPs (e.g. [Alloza et al. \(2020\)](#), [Bijsterbosch et al. \(2017\)](#), [Liu et al. \(2019\)](#), [Lund et al. \(2020\)](#), [Maglanoc et al. \(2019\)](#), [Mandal et al. \(2020\)](#), [Shen et al. \(2018\)](#), or [Vidaurre et al. \(2018\)](#)). Some groups have done some studies relating connectivity with IDPs from other modalities (e.g. [Buchanan et al. \(2020\)](#), [Kernbach et al. \(2018\)](#), or [Shen et al. \(2020\)](#)). Finally, some studies have run their own analysis, using either our processed subject-level functional connectivity maps or our group functional connectivity maps as a reference (e.g. [Anderson et al. \(2018\)](#), [Braga et al. \(2019\)](#), [Cheng et al. \(2018\)](#), [Luo et al. \(2019\)](#), [Moreau et al. \(2020\)](#), [O’Connell et al. \(2020\)](#), [Santana et al. \(2019\)](#), or [Zonneveld et al. \(2019\)](#)).

The study of **brain age** is gaining a lot of momentum in the neuroscience field, and some studies have already benefitted from using the data we generated. For example, these use our IDPs for their research: (e.g. [Anatürk et al. \(2020\)](#), [Cole \(2020\)](#), [de Lange and Cole \(2020\)](#), [Smith et al. \(2019\)](#), and [Smith et al. \(2020\)](#)). A different group of studies just uses the images we generate to evaluate brain age using methods like deep learning, etc. (e.g. [Bashyam et al. \(2020\)](#), [Dafflon et al. \(2020\)](#), [de Lange et al. \(2019a\)](#), [de Lange et al. \(2020a\)](#), [Jonsson et al. \(2019\)](#), [Kaufmann et al. \(2019\)](#), [Kievit et al. \(2018\)](#), [Ning et al. \(2020b\)](#), [Ning et al. \(2020a\)](#), [Peng et al. \(2019\)](#), or [Suzuki et al. \(2019\)](#)).

Finding the **genetic basis of human brain structure and function** has used one of the most interesting combinations of UK Biobank data: genetics and brain imaging. Some of the studies have used our IDPs (e.g. [Barbu et al. \(2019\)](#), [Barbu et al. \(2020\)](#), [Biton et al. \(2019\)](#), [Cheng et al. \(2020\)](#), [Cirulli et al. \(2020\)](#), [Elliott et al. \(2018\)](#), [Georgakis et al. \(2020\)](#), [Persyn et al. \(2020\)](#), [Reus et al. \(2017\)](#), [Satizabal et al. \(2019\)](#), [Sha et al. \(2020\)](#), [Traylor et al. \(2019\)](#), [van Der Lee et al. \(2019\)](#), [Ward et al. \(2019\)](#), [Wei et al. \(2019\)](#), [Wu \(2019\)](#), or [Zhao et al. \(2020\)](#)), while others have just used our processed NIFTI images (e.g. [Anderson et al. \(2020\)](#), [Elvsashagen et al. \(2019\)](#), [Evangelou et al. \(2019\)](#), [Ge et al. \(2018\)](#), [Grasby et al. \(2020\)](#), [Le Guen et al. \(2020\)](#), [Kaufmann et al. \(2018\)](#), [Le Guen et al. \(2019b\)](#), [Neilson et al. \(2019\)](#), [O’Connell et al. \(2020\)](#), [Rodrigue et al. \(2020\)](#), [Soheili-Nezhad et al. \(2020\)](#), [Taquet et al. \(2020\)](#), [van der Meer et al. \(2020\)](#), [Wigmore et al. \(2017\)](#), or [Zhao et al. \(2019b\)](#)).

There have also been very interesting findings regarding **methods research for neu-**

roscience, and our work has proven to be valuable for testing these new methods. In that line, a relatively large number of studies use the NIFTI images generated by us (e.g. Bastiani et al. (2019), Bauermeister et al. (2020), Bowring et al. (2019), Eitel et al. (2019), Glocker et al. (2019), Gong et al. (2020), Hernandez-Fernandez et al. (2019), Kolbeinsson et al. (2019), Ktena et al. (2018), Leming and Suckling (2020), Maximov et al. (2019b), Maximov et al. (2020), Maximov and Westlye (2020), Parker Jones et al. (2018), Pinaya et al. (2020), Pomponio et al. (2020), Rajchl et al. (2018), Stankevičiūtė et al. (2020), Vakli et al. (2020), Wachinger et al. (2020), or Warrington et al. (2019)). Other methods-development studies use the IDPs (e.g. Abrol et al. (2020), Arslan et al. (2018), Davenport and Nichols (2020), He et al. (2018), Pervaiz et al. (2020), Schulz et al. (2019), Dai et al. (2019), or Zhou et al. (2019)).

Finally some parts of our most recent work on confounds has already been used in Elliott et al. (2018), Smith et al. (2019), Smith et al. (2020), and Wiberg et al. (2019).

5.3 Future work: Pipeline

The processing pipeline described in Chapter 2 was developed in 2016, and keeping it robust, consistent and unchanged¹ was one of our main goals. Nevertheless, the field has advanced since then and updating it to a new version (while also re-processing all the data with this version) will be very valuable. Over the next year or two we plan to investigate the following improvements and expansions for a "v2" UK Biobank processing pipeline:

5.3.1 Strategic changes in the pipeline

- Make the pipeline highly configurable from the command line or with a configuration file (allow for running just certain sections, exclude others, avoid GDC, set parameters, etc. Default would be "default UK Biobank").
- Minimise the number of cluster jobs per subject-batch. 2 options:
 - Horizontal minimisation: Join together all jobs of the same "block" per subject, like joining all "non-GPU jobs for subject X in one job".
 - Vertical minimisation: Create 1 job per script for all subjects and make them co-dependent (empty jobs when a subject does not have the modality).
- Adopt BIDS (Gorgolewski et al., 2016) and BIDS-derivatives (Markiewicz et al.,

¹Or rather, backwards compatible, meaning that we have only added new outputs, and have not changed existing ones.

2020) as the standard.

- Still, include a converter to the original directory structure.
- Use Python whenever possible (avoid bash):
 - Use fslpython as the general environment and fslpy when possible (MacCarthy, 2017).
 - Evaluate internally (WIN - FMRIB) the use of nipype (Gorgolewski et al., 2011).
 - Replace the python installer with something more pythonic (create a python package).
- Distribute the pipeline with a neurodocker spec or similar container as a BIDS-app (Gorgolewski et al., 2017).
- Switch to a single multimodal volumetric registration per subject - at least covering T1, T2-FLAIR and dMRI eigenvectors (and e.g. FA). MMORF (Lange et al., 2020) is now almost ready for this role and performing extremely well.
- Consider what to do about bias field correction for all modalities.
 - Include the total bias calculated from T1_not_norm and apply all the transformations to have e.g. B0 without bias field.
- Include ASL processing (Chappell et al., 2008) in the pipeline (UK Biobank does not yet have ASL, but the pipeline would be more general this way, and UKB may include this for later subjects/timepoints).
- Organise brainstorm sessions including external researchers
 - Consult with BIDS people to standardise as much as possible the BIDS derivatives.
 - Consult about what can be improved for each modality.
- Think about ways to ease the testing of each module.
- Investigate adding longitudinal (two-time-point) analyses (e.g. SIENA, longitudinal-FS, etc).

5.3.2 General changes in the pipeline processing

- Include more information in the log files about the FSL version, the python version, the libraries, etc².
- Use fsl_deface instead of the complex struct_init script (generate faces in 3D with

²Suggestion: reprozip <https://www.reprozip.org/>.

FSLEyes as a QC feature to check).

- Consider use of `fsl_deface` with SWI, not just T1 and T2 FLAIR
- Investigate whether to add motion confound regressors into the tfMRI GLM.
- Switch to a single integrated modified version of `gradunwarp`.
- Get vein maps (Division T1/T2), e.g. for FIX.

5.3.3 Initial / final Processing

- Switch to the latest version of `dcm2niix` ([Li et al., 2016](#)). Make sure to use properly all the new functionalities: Include `ImaAbsTablePosition` and `TablePositionOrigin` from DICOMs.
- Check every feature from the DICOMs (echo time, pixel size, etc) for all subjects and create a dictionary of what is acceptable.
- Include a general check at the end that everything completed correctly and warns if not

5.3.4 T1

- Correct brain mask by adding the overlapping with the defacing mask. ensure that no voxels are missing.
- Brain extraction:
 - For SIENAX, adapt the skull-estimation output from BET (will need a code change in BET) so that bet can be fed the good brain extraction already generated by the pipeline and not rely on BET starting again from scratch for brain extraction (which does not work very well on UKB data). I.e. we will feed in T1 and T1_brain into BET, when asking for skull output.
 - Investigate - as a new processing output - using the dMRI “nodif” image (in place of a T2) for full betsurf analysis, so that we can get the inner skull surface and hence a better ICV estimate.
- Consider multi-modality methods like T1 + FLAIR versus T1 in FAST, or using T1 + FLAIR + dMRI versus T1 + FLAIR in BIANCA.

5.3.5 T2 FLAIR

- Update to the latest BIANCA ([Griffanti et al., 2016](#); [Sundaresan et al., 2019a,b](#)) or possibly an improved approach based on deep learning. Train with more subjects.

- Add regional BIANCA IDPs.
- Check what happens with extreme WMH subjects and whether they should be unusable for some modalities (fMRI / dMRI). Check also what happens with dMRI in these subjects.

5.3.6 swMRI

- Include QSM processing.
- Follow the same naming standard in SWI as in T1.
- Verify scaling of the filtered_phase in SWI.

5.3.7 FreeSurfer

- Evaluate the use of FastSurfer ([Henschel et al., 2020](#)).
- Run HCP-like processing. Obtain activation and connectivity maps on the surfaces (CIFTI).
 - Decide whether we should do as HCP does and go straight to 2mm MNI space for 4D preprocessed fMRI data.
 - Perform surface-based registration with Multimodal Surface Matching ([Robinson et al., 2014](#)) using morphometric measures such as cortical thickness or cortical surface area, mapping resting fMRI and dMRI onto the cortical surface.
 - As the resolution, quality and quantity (e.g., number of fMRI timepoints) in UK Biobank data cannot match that in the HCP (with its 4 hours of scanning per subject) each of these steps will involve careful evaluation and potential reworking in the context of our data. For example, the projection of the lower-resolution fMRI data onto the cortical surface vertices, including the HCP approach of removing “noise” voxels, will need careful optimisation.

5.3.8 Fieldmap

- Check whether it is possible to work without fieldmaps to avoid losing fMRIs because of missing dMRI.

5.3.9 dMRI

- Use the improved mask in autoPtx (this will happen once we switch to MMORF ([Lange et al., 2020](#))).
- Use the newest eddy and add the motion reports (and any other report generated by eddy) into the IDPs. Make sure to use the BIDs-compliant options.
- Use the eddy_rotated_bvec in the whole diffusion pipeline.
- Evaluate recommendations from ([Maximov et al., 2019a](#)).
 - Denoising using the PCA method from ([Veraart et al., 2016](#)).
 - Correct Gibbs ringing.
 - Consider bias field correction.
- Update to XTRACT ([Warrington et al., 2020](#)) instead of autoPtx .
- Confirm new registration works in the middle cerebellar peduncle (currently poor).
- Consider the latest NODDI modelling (CUDIMOT) ([Hernandez-Fernandez et al., 2019](#)) and other new methods for microstructural modelling.

5.3.10 fMRI

- Consider whether bias field correction on fMRI data may improve motion correction or cross-subject statistics.
- Investigate using FREDDY (eddy for fMRI) for motion correction.
- Improve the brain masking for fMRI. In some cases, it is too restrictive, and in some others, it shows activation out of the brain.
- Develop a test to auto-detect whether the SBRef is correct, based on it having high enough within-brain contrast. Work out what to do when SBRef is not good enough.
- Consider replacing dual-regression with “PROFUMO-predefined” ([Harrison et al., 2015, 2020](#)).

5.4 Future work: QC

There are plenty of possible ways in which the QC processing may be improved in future instances. Some of these possibilities include:

- Create a warning file if any T1 IDPs are NaN.
- Generate normative models for key features (mean intensity, 5th and 95th percentile) and define thresholds to find outliers for each modality before processing them.

- Debate including QC features and QC labels in the dataset. Also, QC features out of the QC tool (COG, intersection between the defacing and the brain mask, etc.).
- Expand, improve, and train autoQC. Revalidate autoQC.
 - Include in the autoQC more region-driven features.
- Reprocess subjects with category 40 → There is Normalised T1, but there is no NOT-norm T1.
- Discuss whether it would be even better to discard “incomplete” (e.g. incomplete FOV) T1-related images before running the file manager logic.
- Expand basic, automatable QC for all modalities.

5.5 Future work: Confounds

Even though the confound work is very recent and thorough, there are some paths that may still be explored in the near future. For example, we may want to explore the relationship of the described confounds with non-imaging confounds such as BMI or blood pressure. How these generally used confounds relate to the proposed confounds, the IDPs and the correlations between IDPs and non-IDPs may be of great interest.

Another interesting feature related to the studied confounds would be to study the relationship between head size (from SIENAX), brain size, and intracranial volume (from FreeSurfer, which probably partially reflects brain size and not just head size). This study may involve studying the genetic component of this relationship.

On a different note, even though it has been argued that unconfounding may be a counter-productive process in a prediction-oriented Machine Learning setting ([Chyzhyk et al., 2018](#)), it may be useful in a causal inference setting. Therefore, studying the QC features developed and explained in Section 3.3.1 from a causal point of view, including several confounds in the process may give better insights into what causes an image to be deemed unusable (bad quality) and may help in future QC research.

Appendix

A.1 Motion and SNR in the imaging modalities

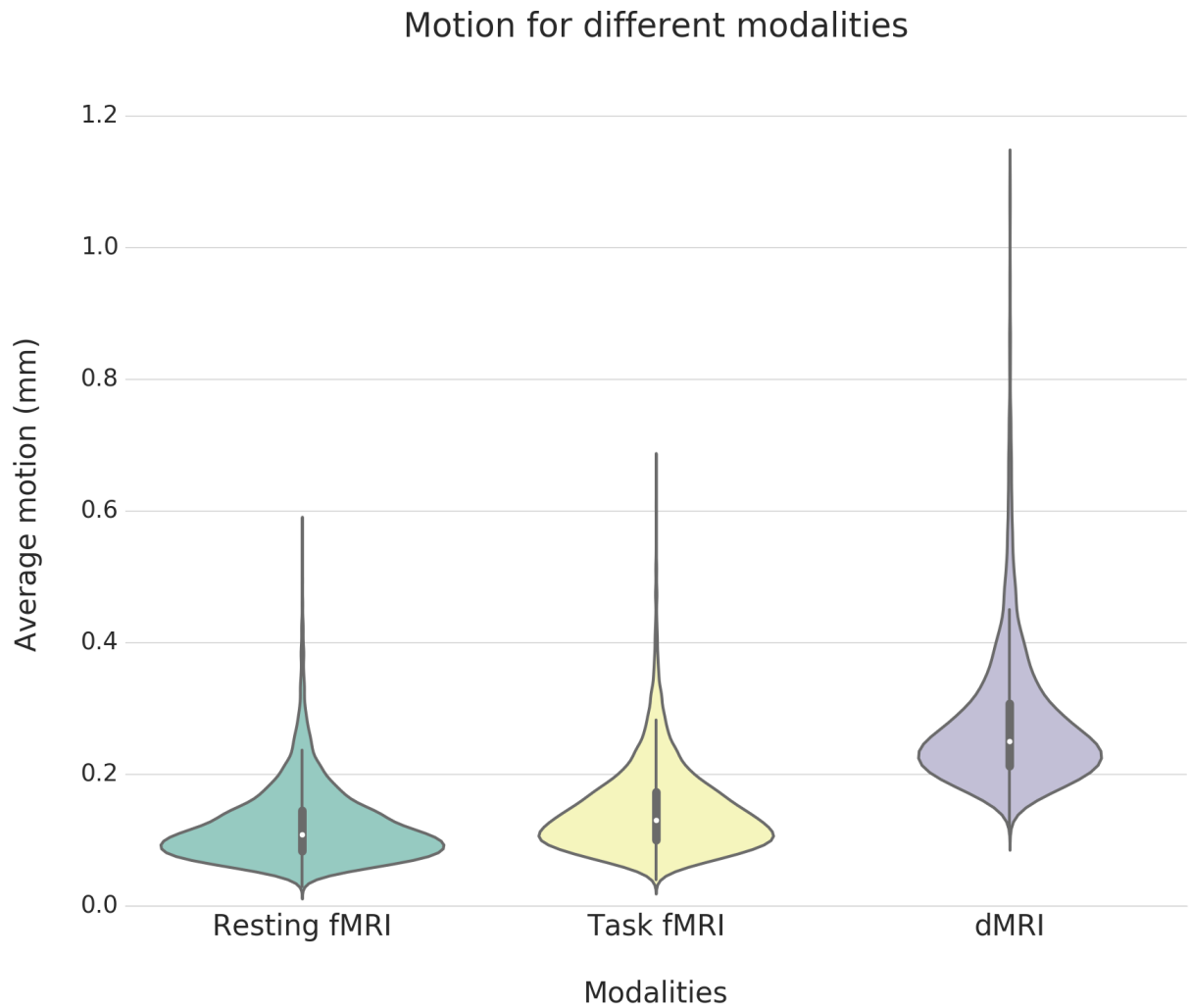


Figure A.1: Motion estimation for the temporal modalities. The motion per modality is calculated using the motion of each timepoint relative to the previous one, and averaging the motion displacement across the brain and then across timepoints. For dMRI, motion is calculated without the translation in the AP encoding direction (Y), as that component is highly affected by eddy distortions. Movement is higher in dMRI because, assuming that much of the true movement is slow over time in comparison to the TR, volume-to-volume movement should be approximately proportional to TR, which is four times larger. Extreme outliers (points further than $mean \pm (8 * \sigma)$) have been removed for ease of visualisation.

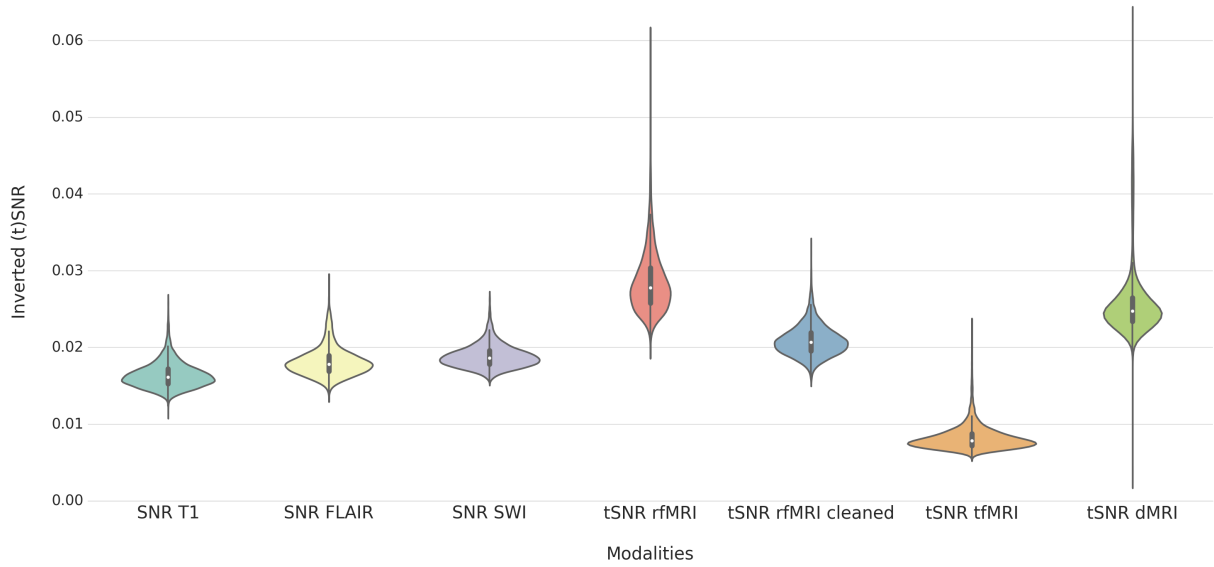


Figure A.2: Inverted SNR for T1, T2 FLAIR and SWI (magnitude, first echo) and inverted tSNR for Resting fMRI, Resting fMRI (after removing noise components using FIX), Task fMRI (processing includes spatial smoothing, as explained in Section 2.2.9 of the main text) and dMRI (calculated using all b=0 images in the AP direction). Extreme outliers (points further than: $mean \pm (8 * \sigma)$) have been removed for ease of visualisation.

A.2 Evaluation of change in shimming

The new auto-shimming approach was put in place with a reduced shimming field-of-view and fewer shim iterations. Some tests, described below, were performed in order to determine whether these changes resulted in any disruption to the data.

Shimming is the process of readjusting the homogeneity of a magnetic field during the scanning. In order to reduce the total acquisition time of the protocol¹, the number of shimming iterations was reduced to one, performed at the beginning of the acquisition protocol. Two features relating to the fMRI EPI data, “warping extent” and “drop-out” were checked and compared in 100 subjects from before the change and 100 after the change. “Warping extent” refers to the distortion correction displacement (distance), in voxels, represented as a displacement map estimated for fMRI images while “drop-out” refers to the intensity loss artefact (for example in temporal lobes) in fMRI EPI images. It is worth noting that we were not looking for these values to improve, but rather to achieve similar levels of warping/drop-out despite the reduced time spent on shimming.

- “Warping extent”: Figure A.3.A shows that the distortion correction estimated for EPI images was not substantially affected by the shimming change (the histograms

¹Time saved with this change was 1 minute 28 seconds.

show a very slight reduction in distortion with the latter shimming method).

- “Drop-out”: We defined drop-out regions as being brain regions where intensity (averaged across subjects) falls to less than 20% of the global mean intensity. Figure A.3 (panels B and C) shows the mean signal drop-out (focused on the temporal lobes) for 100 subjects before and after the shimming change, with only minor differences apparent. There were no significant changes in mean intensity in drop-out regions after correcting for multiple comparisons across space. At a more liberal testing threshold ($P < 0.01$, two-tailed t-test, uncorrected for multiple comparisons), there was a small medio-frontal area of change in signal level ($t_{200} = 3.4$); this showed less drop-out with the latter (customised) shimming approach as can be seen in Figure A.3.D. This is also confirmed with histograms in Figure A.3.E, which shows a strong similarity of the intensity distributions in rfMRI before and after the change.

The primary reason that the reduction in EPI warping and drop-out does not get worse here with fewer shimming iterations is that the volume the shim is optimised over is now reduced and (accordingly) the shim data resolution is improved.

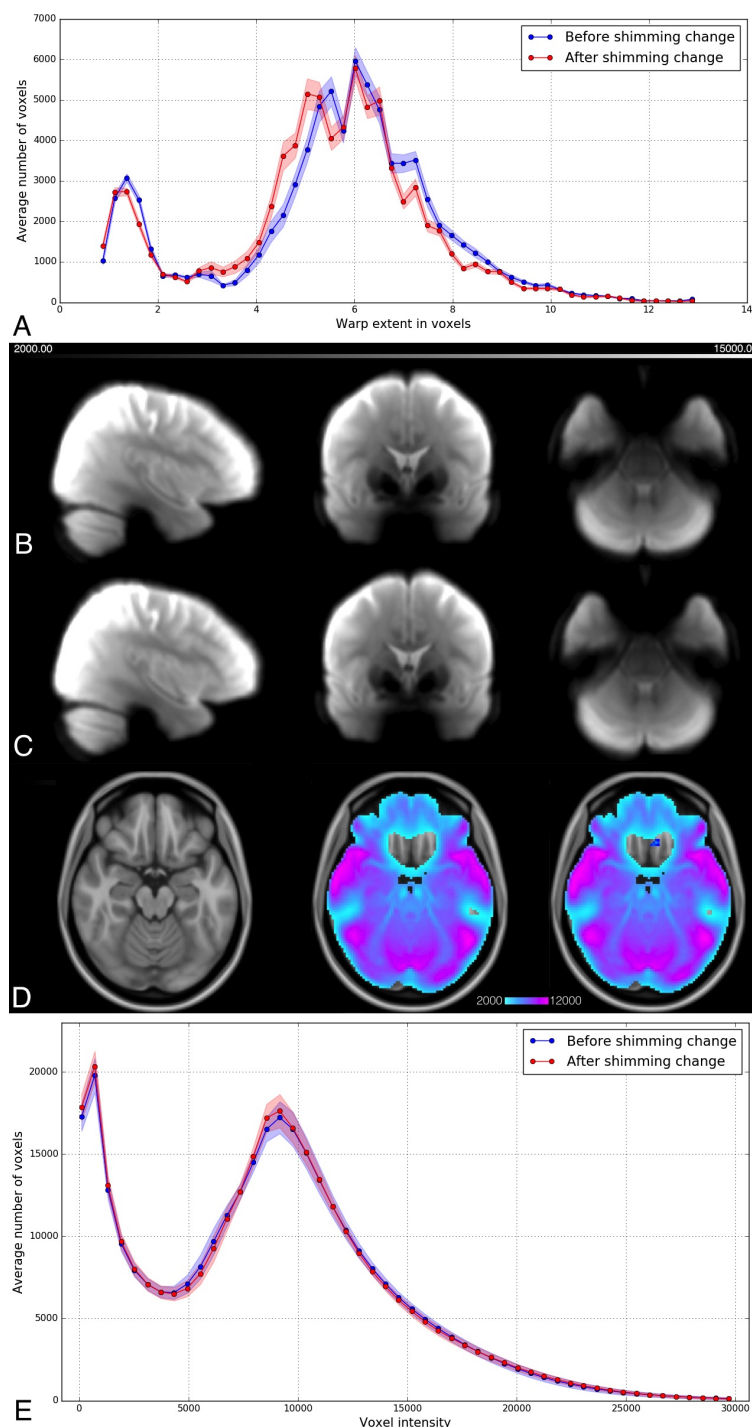


Figure A.3: Tests to verify that the change in shimming did not affect the data significantly. *A*: Histogram across voxels of the undistortion warp extent before and after the shimming change (Ribbon = standard deviation) averaged across subjects. *B*: Mean drop-out for 100 subjects before shimming change and (*C*) after shimming change. *D left*: MNI152 T1 template. *D middle*: Mean fMRI, original shimming thresholded at 20% global mean. *D right*: Mean fMRI, customised shimming thresholded at 20% global mean. Difference between two shimming methods in dark blue ($P < 0.01$ uncorrected). *E*: Histogram across voxels of the mean intensity of EPI images across 100 subjects (Average of rfMRI) (Ribbon = standard deviation).

A.3 Acquisition and protocol features

Site	Number of subjects
Stockport	24,744
Newcastle	9,842
Reading	5,111

Table A.1: As read from DICOM headers

Processing batch	Number of subjects
1	5,733
2	4,217
3	4,570
4	6,885
5	2,070
6	16,222

Table A.2

CMRR Version	Number of subjects
vd13a-R010-r-5b5bc96-Nov-21-2013-02	92
vd13a-master-r-6823e50-Jun-7-2014-04	1,384
vd13a-master-r-df0bdb6-Nov-19-2014-20	2,069
vd13a-R012-r-b34510a-May-18-2015-18	2,939
vd13a-master-r-c0cb646-Dec-15-2015-03	2,704
vd13a-master-r-bd40f27-Jul-21-2016-02	3,076
vd13a-master-r-c3fbc7c-Mar-3-2017-23	4,552
vd13a-ukbb-r-804c5139-Jul-18-2017-23	22,881

Table A.3: As read from DICOM headers

Date of Service Pack change	Number of subjects
31/01/2017	11,759
Up to date	27,938

Table A.4: As read from DICOM headers

Protocol version (as read from DICOMs)	Number of subjects
UK Biobank v2.0	554
UK Biobank v3.0	249
UK Biobank v4.0	673
UK Biobank v5.0	7,712
UK Biobank v6.0	3,813
UKBiobank_Head v6.0	21,585
UK_Biobank^ Head (Site 3)	5,111

Table A.5: As read from DICOM headers

Site	Date of Ramp-down event	Number of subjects
1	Before 22/07/2017	14,187
	Before 03/01/2018	2,179
	Before 24/02/2018	633
	Before 2019/08/29	7,245
	After 2019/08/29	500
Other sites and no events		14,953

Table A.6: As read from DICOM headers

Site	Date of Head Coil change	Number of subjects
1	Before 19/06/2014	154
	Before 26/11/2016	10,817
	Before 2019/01/11	10,515
	After 2019/01/11	3,258
2	Before 20/02/2018	2,349
	After 20/02/2018	7,493
Other sites and no events		5,111

Table A.7: As read from DICOM headers

Site	Date of Cold Head change	Number of subjects
1	Before 29/08/2019	24,244
	After 29/08/2019	500
2	Before 25/10/2019	9,786
	After 25/10/2019	56
	No coil change	925
Other sites and no events		5,111

Table A.8: As read from DICOM headers

Site	Date of Miscellaneous scanner event	Number of subjects
1	Before 16/04/2016 (E-shim module replaced)	7,687
	Before 07/12/2018 (maintenance)	13,461
	Before 05/07/2019 (maintenance)	2,725
	After 05/07/2019 (present)	871
2	Before 25/11/2017 (maintenance)	1,682
	Before 28/12/2018 (maintenance)	4,455
	After 28/12/2018 (present)	3,705
Other sites and no events		5,111

Table A.9: As read from DICOM headers

Flipped SWI dimensions	Number of subjects
False	32,760
True	3,355
No SWI	3,582

Table A.10

FreeSurfer	Number of subjects
With T1 + T2 Flair	38,173
With T1	1,301
No FreeSurfer	223

Table A.11

Latest Eddy	Number of subjects
Original	33,232
Latest	6,465

Table A.12

Echo time for rfMRI (ms)	Number of subjects
39.0	14,780
42.4	24,805
Other values	2
No usable data	110

Table A.13

Echo time for tfMRI (ms)	Number of subjects
39.0	13,699
42.4	20,958
Other values	3
No usable data	5,037

Table A.14

Intensity scaling factor for T1	Number of subjects
1	552
6	39,145

Table A.15

Intensity scaling factor for T2 FLAIR	Number of subjects
1	538
2	3
5	8,659
6	29,767
No usable data	740

Table A.16

Intensity scaling factor for swMRI	Number of subjects
1	11,996
2	24,249
6	2
No usable data	3,450

Table A.17

Intensity scaling factor for dMRI	Number of subjects
1	12,100
2	26,182
6	2
No usable data	1,413

Table A.18

Intensity scaling factor for rfMRI	Number of subjects
1	518
2	39,068
6	1
No usable data	110

Table A.19

Intensity scaling factor for tfMRI	Number of subjects
1	511
2	34,142
6	7
No usable data	5,037

Table A.20

A.4 Raw confound processing algorithms

Algorithm for processing of quantitative confounds

Quantitative confounds processing

```

1 # Requires: confounds, list_sites, list_ind_subj_by_site, numb_of_conf
2
3 # Demedian globally each column and standardise with the mad
4 for i = 1 : numb_of_conf
5     confounds[:,i] = confounds[:,i] - median(confounds[:,i])
6
7     # Normalise by Median Absolute Deviation
8     madV = mad(confounds[:,i]) * 1.48
9
10    # In case madV is 0, use the std
11    if madV < minimum_possible_values
12        madV = std(confounds[:,i])
13    end
14    confounds(:,i) = confounds[:,i] / madV
15 end
16
17 # Split the columns: 1 per site. 0-padding by default.
18 finalConfs = zeros(length(ALL_IDs), numb_of_sites * numb_of_conf)
19
20 # For each Site
21 for site in list_sites do
22     # For each variable (column) that is received in "values"
23     for j = 1: numb_of_conf
24         V = confounds[list_ind_subj_by_site[site], j]
25         medTP = median(V)
26
27         # Outlier removal
28         V[find(V > 8)] = NaN
29         V[find(V <-8)] = NaN
30
31         # Replace all NaNs (also outliers) with the median
32         V[isnan(V)] = medTP
33
34         # Normalise by site
35         V = normalise(V)
36         final_conf[list_ind_subj_by_site[site], (numb_of_conf*(i-1) + j)] = V
37     end
38 end

```

The script the performs these operations is [\[LINK\]](#).

Algorithm for processing of qualitative confounds

Qualitative confounds processing

```

1  # Requires: confound, list_sites, list_ind_subj_by_site, numb_IDPs
2
3  # For each Site
4  for site in list_sites do
5      numb_new_cols = 0
6      ind_subj_by_site = list_of_ind_subj_by_site[site]
7      numb_subj = length(ind_subj_by_site)
8      different_values = unique(confound[ind_subj_by_site])
9      different_values = different_values[~isnan(different_values)]
10
11     numb_of_different_values = length(different_values)
12
13     # If all elements for this site have the
14     # same value, we do not generate a column.
15     # Subjects from this site will have 0 in
16     # the columns for the other sites.
17     if numb_different_values > 1
18         numb_new_cols = numb_different_values-1
19         new_conf = zeros(numb_IDPs, numb_different_values-1)
20
21         # Subjects (this site) with 1st value get a -1 in all new columns
22         indices_for_value = find(confound == different_values[0])
23         index = intersect(indices_for_value, subj_by_site)
24         index_final = indices_for_value[index]
25
26         new_conf[index_final,:] = -1
27
28         # Each new value gets a new column. Subjects (of this Site) with
29         # this value have 1 in this column. All other subjects have a 0.
30         for j = 2:numDiffValues
31             indices_for_value = find(confound == different_values[j])
32             index = intersect(indices_for_value, subj_by_site)
33             index_final = indices_for_value[index]
34
35             new_conf[index_final,j-1] = 1
36
37             ind_not_zero = find(new_conf[:,j-1] ~= 0)
38             new_conf[ind_not_zero,j-1] = normalise(new_conf[ind_not_zero,j-1])
39         end
40
41         # Add the new columns generated for this
42         # site to the final set of columns
43         final_conf = [final_conf, new_conf]
44     end
45 end

```

The script the performs these operations is [\[LINK\]](#).

A.5 Non-additive terms (NATs) processing algorithm

Algorithm for processing of NATs

NAT processing

```

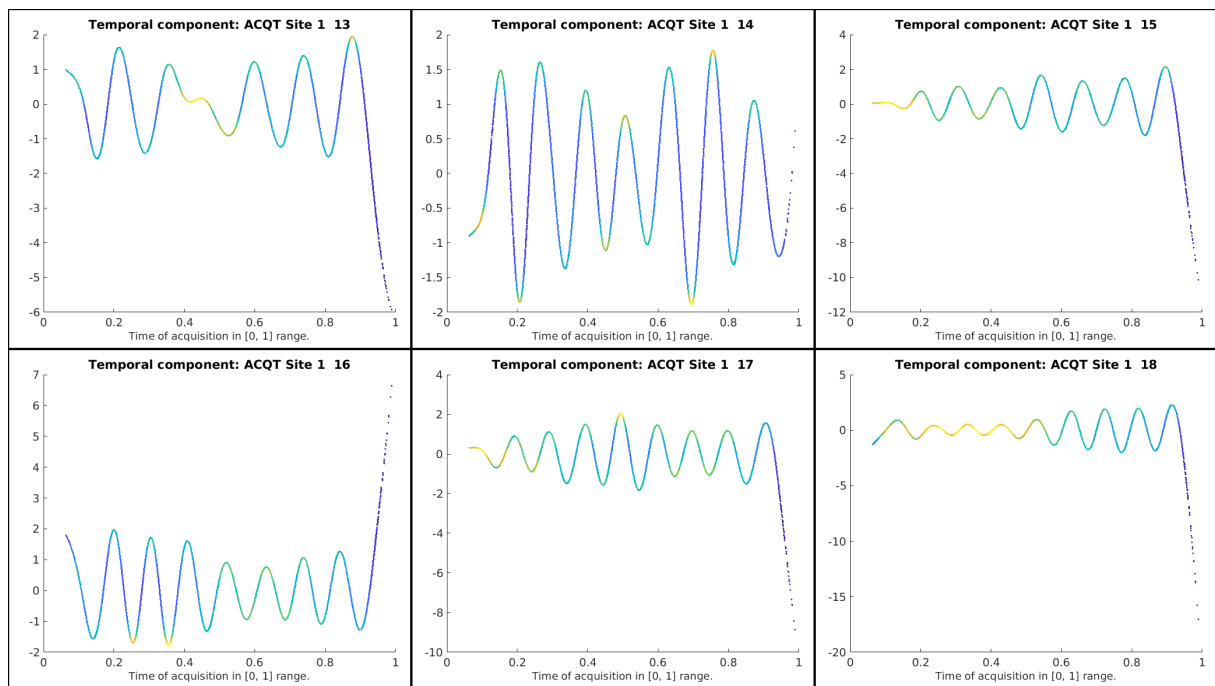
1 # Requires: confounds\_to\_check, IDPs
2
3 #  $J = I + aA + bA^2 + cAI + \text{noise}$ 
4 # ( $A = \text{confound} / I = \text{TrueIDP} / J = \text{MeasuredIDP}$ )
5 # For each confound to check
6 for A in confounds_to_check
7
8     % For each IDP
9     for J in IDPs
10         AA = A.\^2;
11         X=[A AA];
12         [a b] = pinv(X)*J;
13         # 1st pass at estimating I, ignoring interact. term cAI
14         I1 = J - X * [a b];
15
16         # Iterate 20 times, including interaction term cAI
17         for i = 1:20
18             AI = A .* I1;
19             X = [A AA AI];
20             [a b c] = pinv(X) * J;
21             I2 = J - X * [a b c];
22             # Only do fractional update so we don't oscillate
23             I1 = (I1 + I2) / 2;
24         end
25     end
26 end

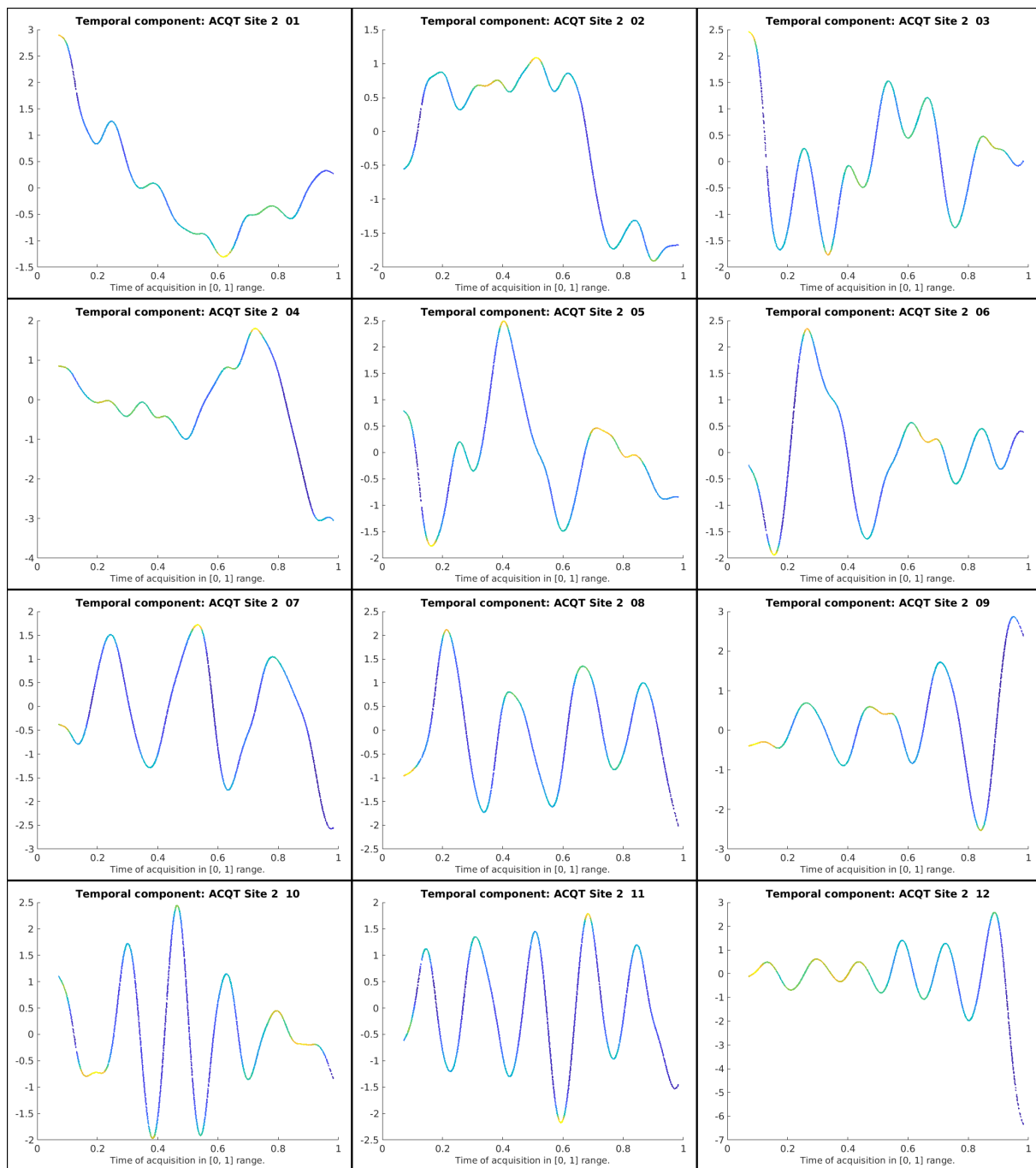
```

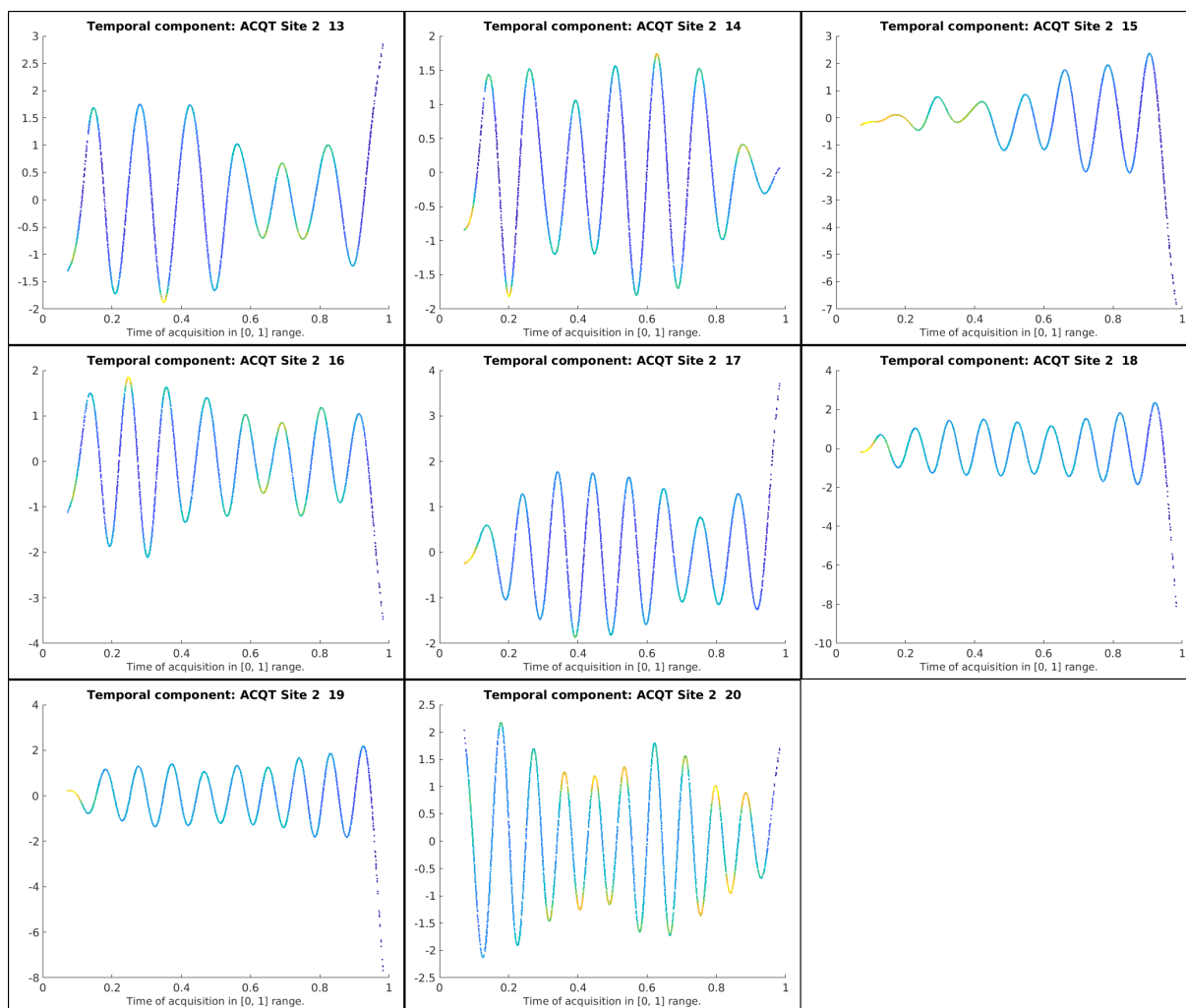
The exact script the performs these operations is in [\[LINK\]](#)

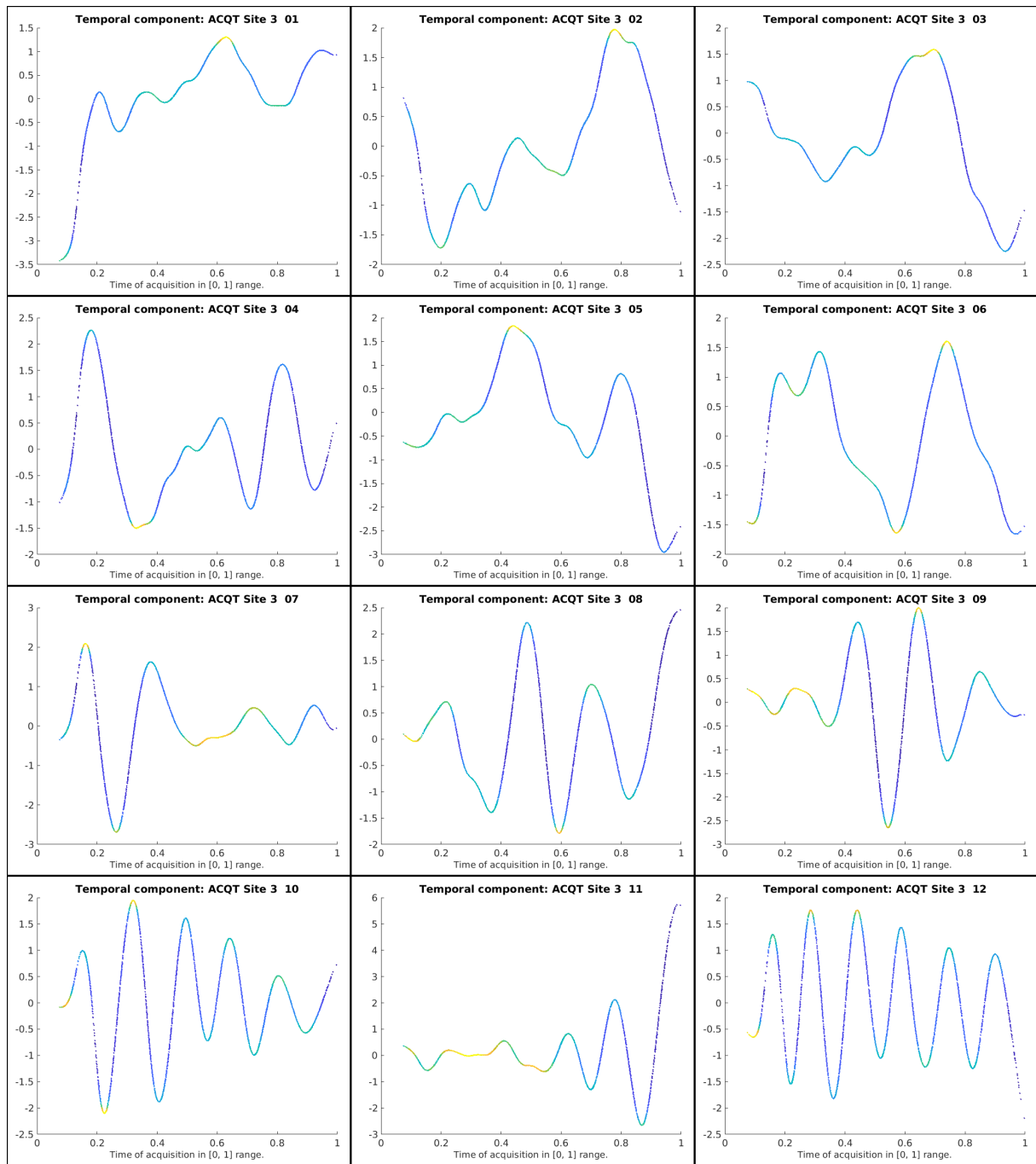
A.6 Temporal components

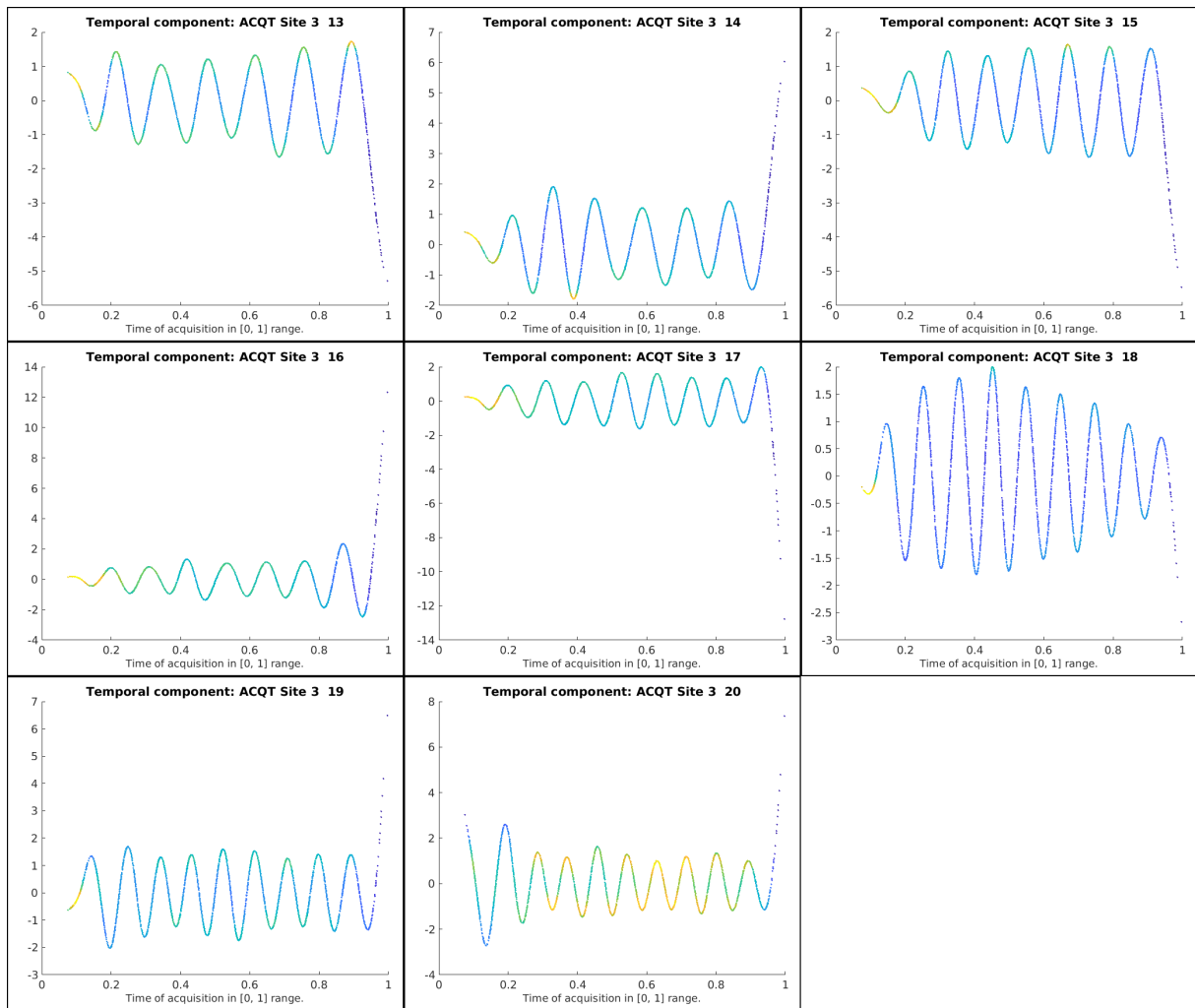






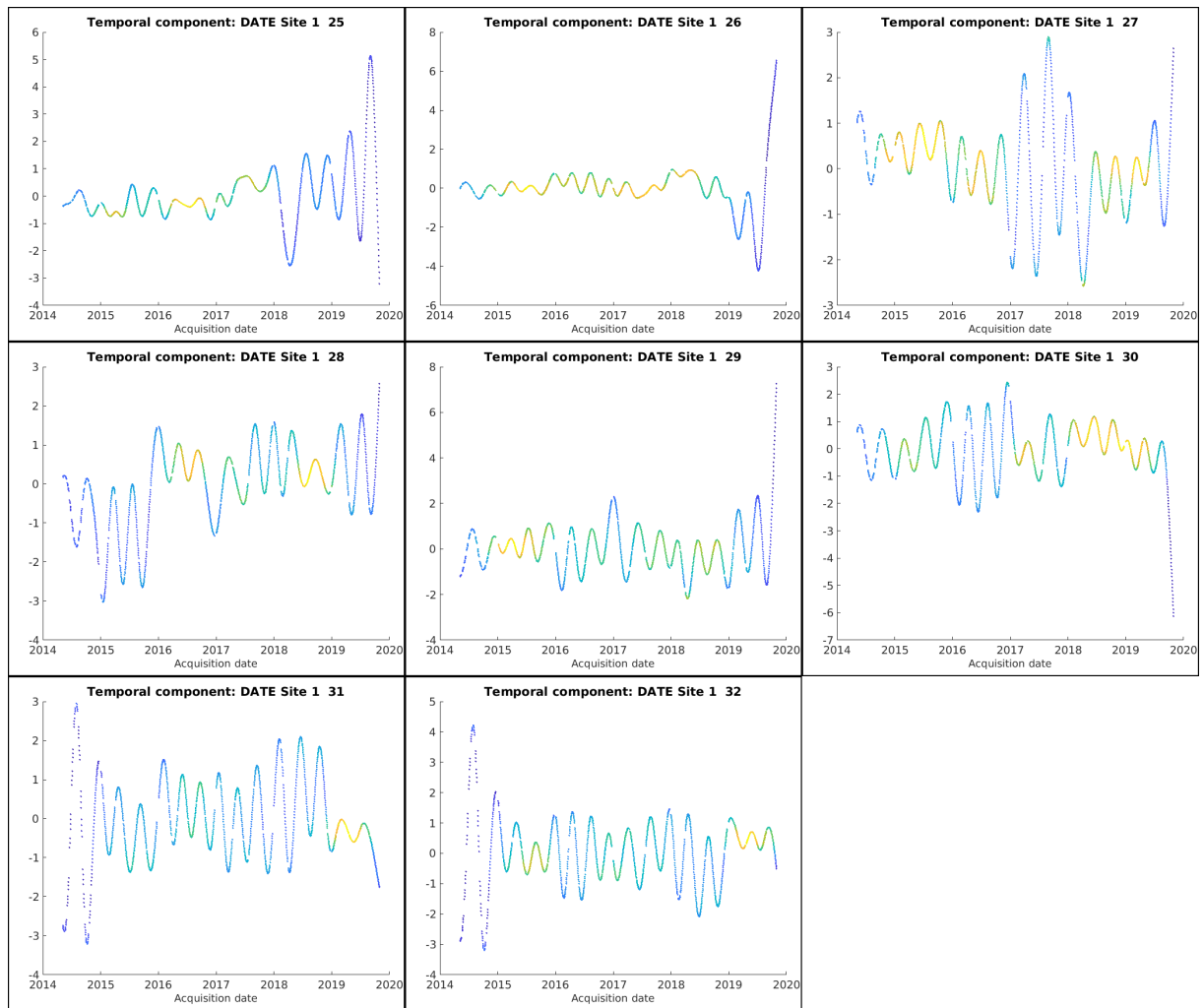


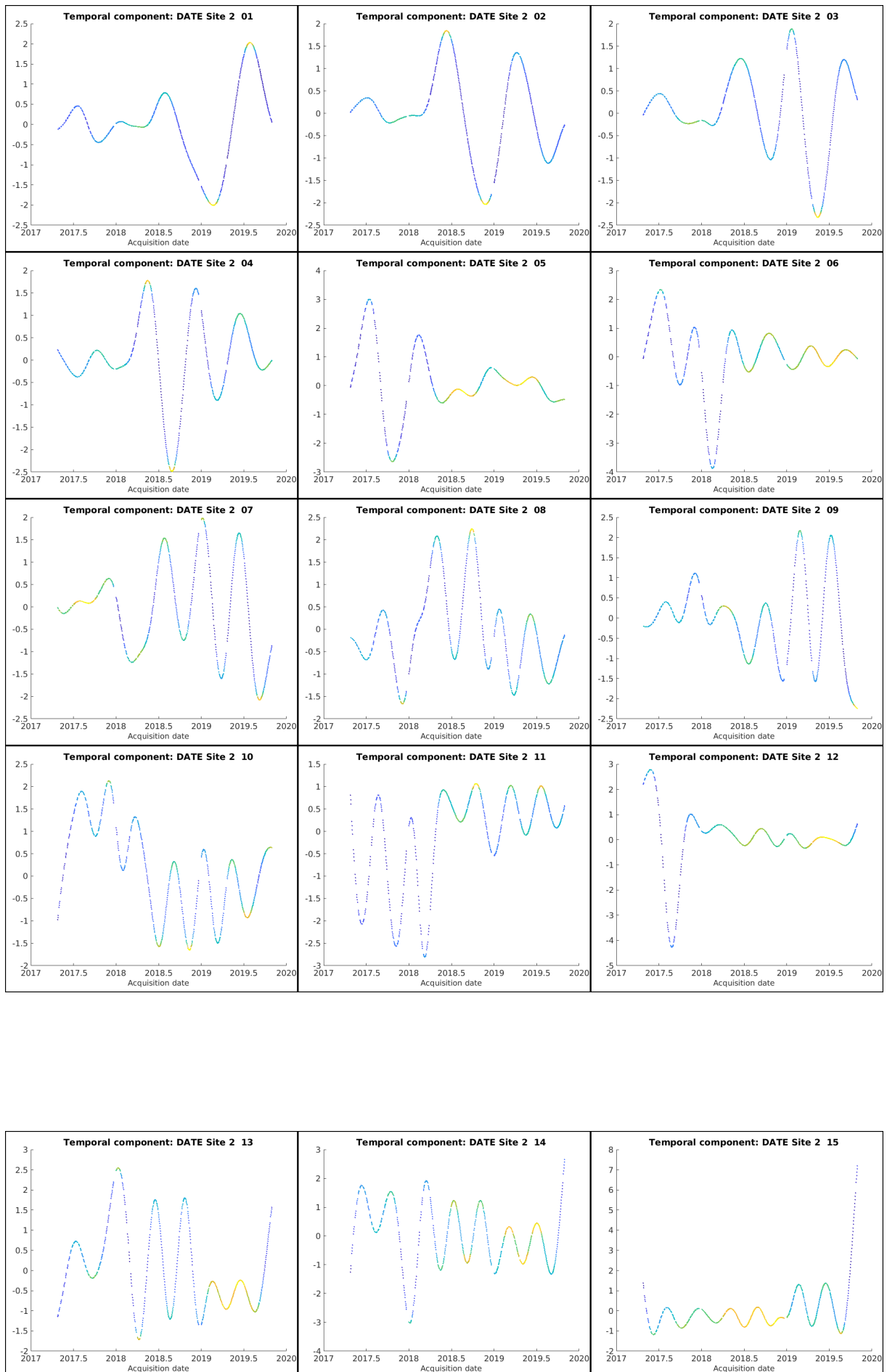


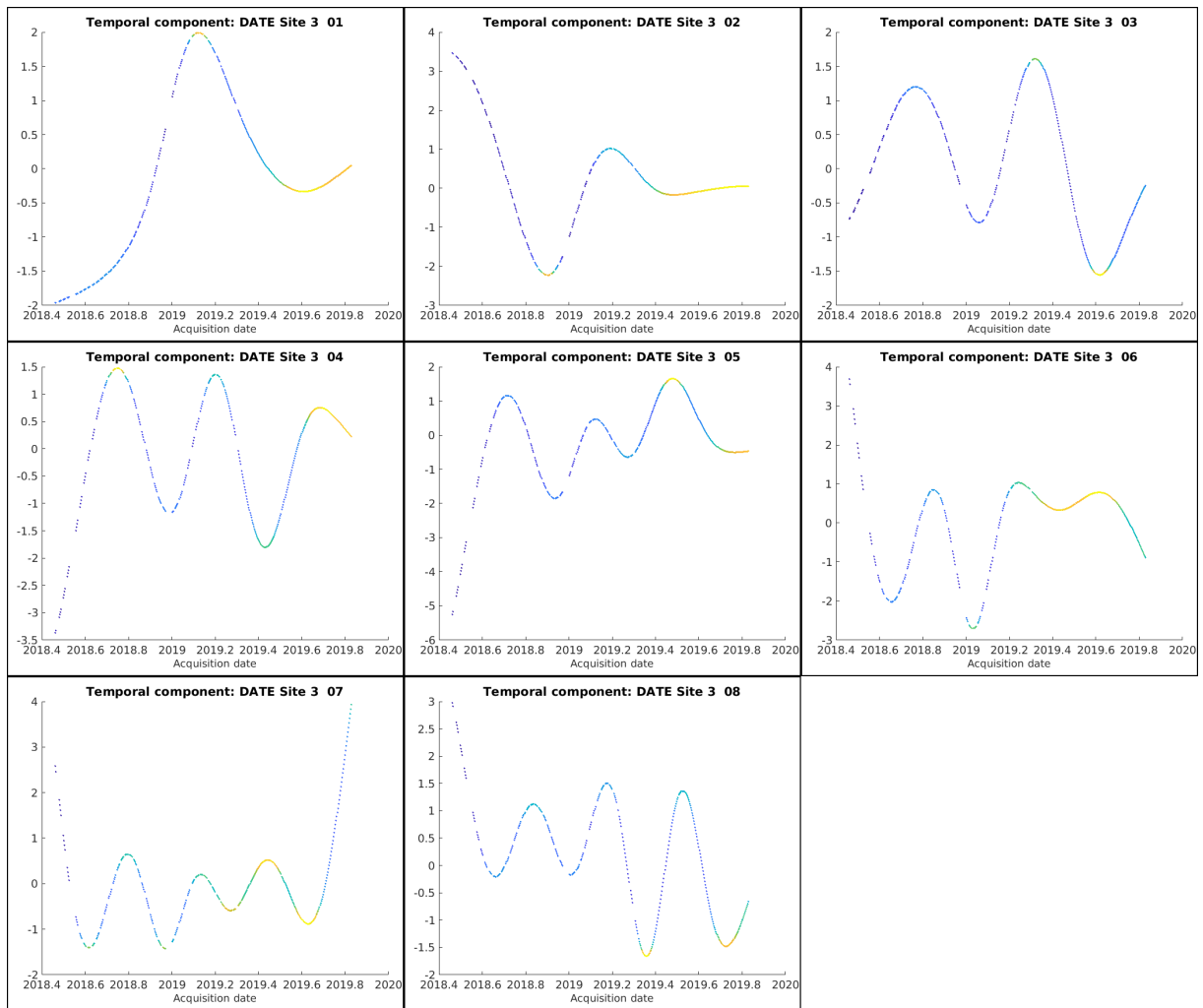




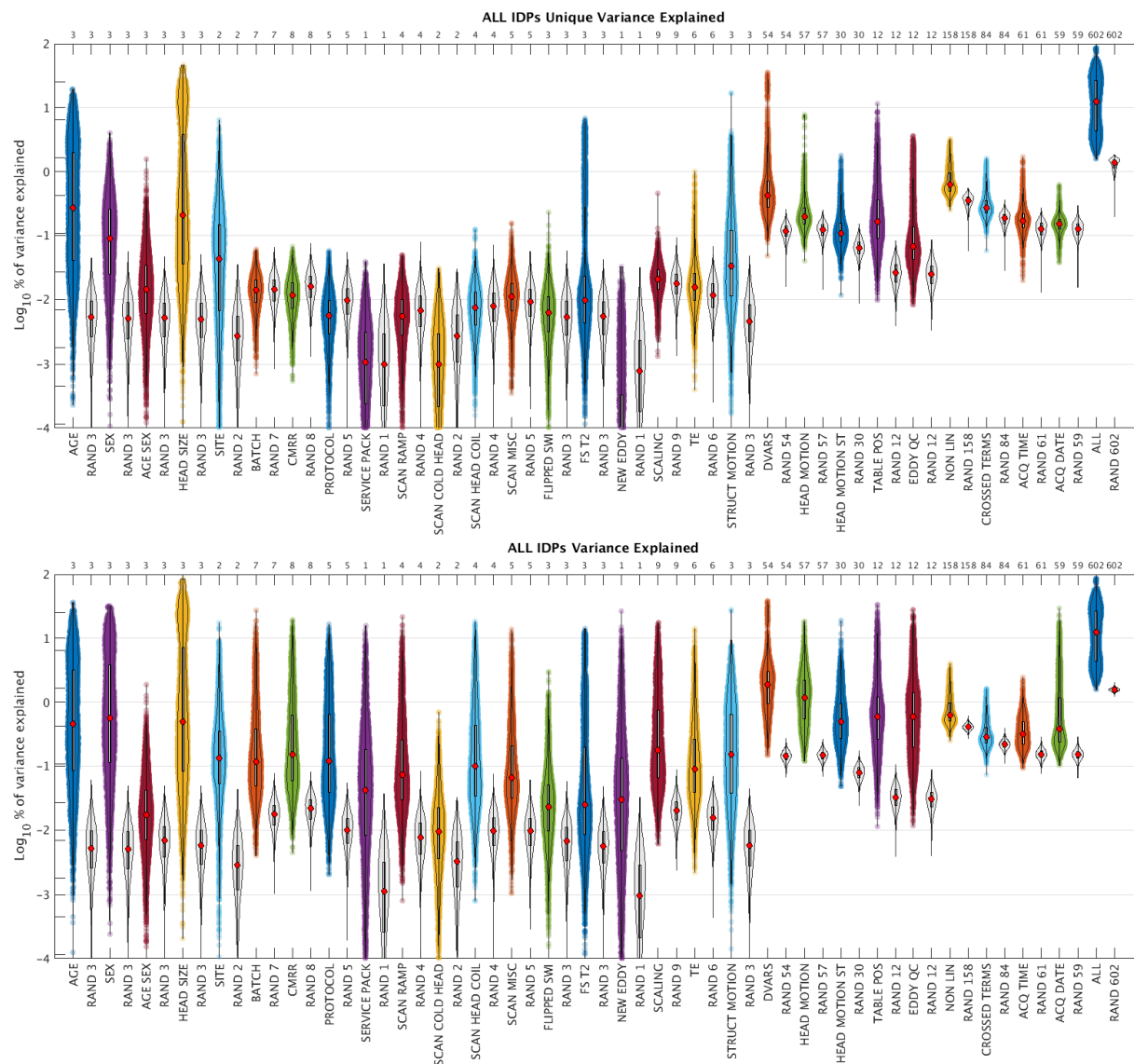


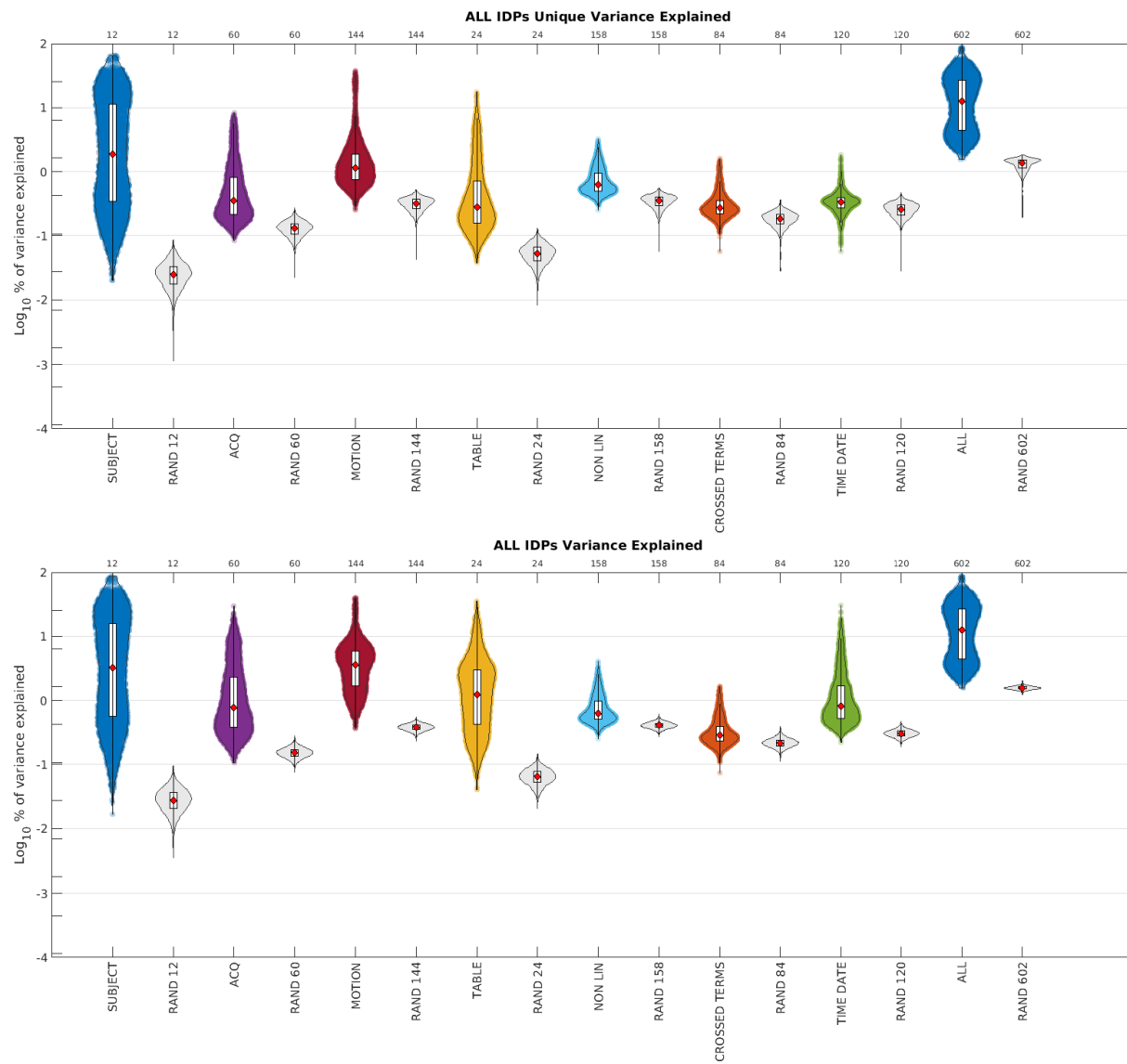


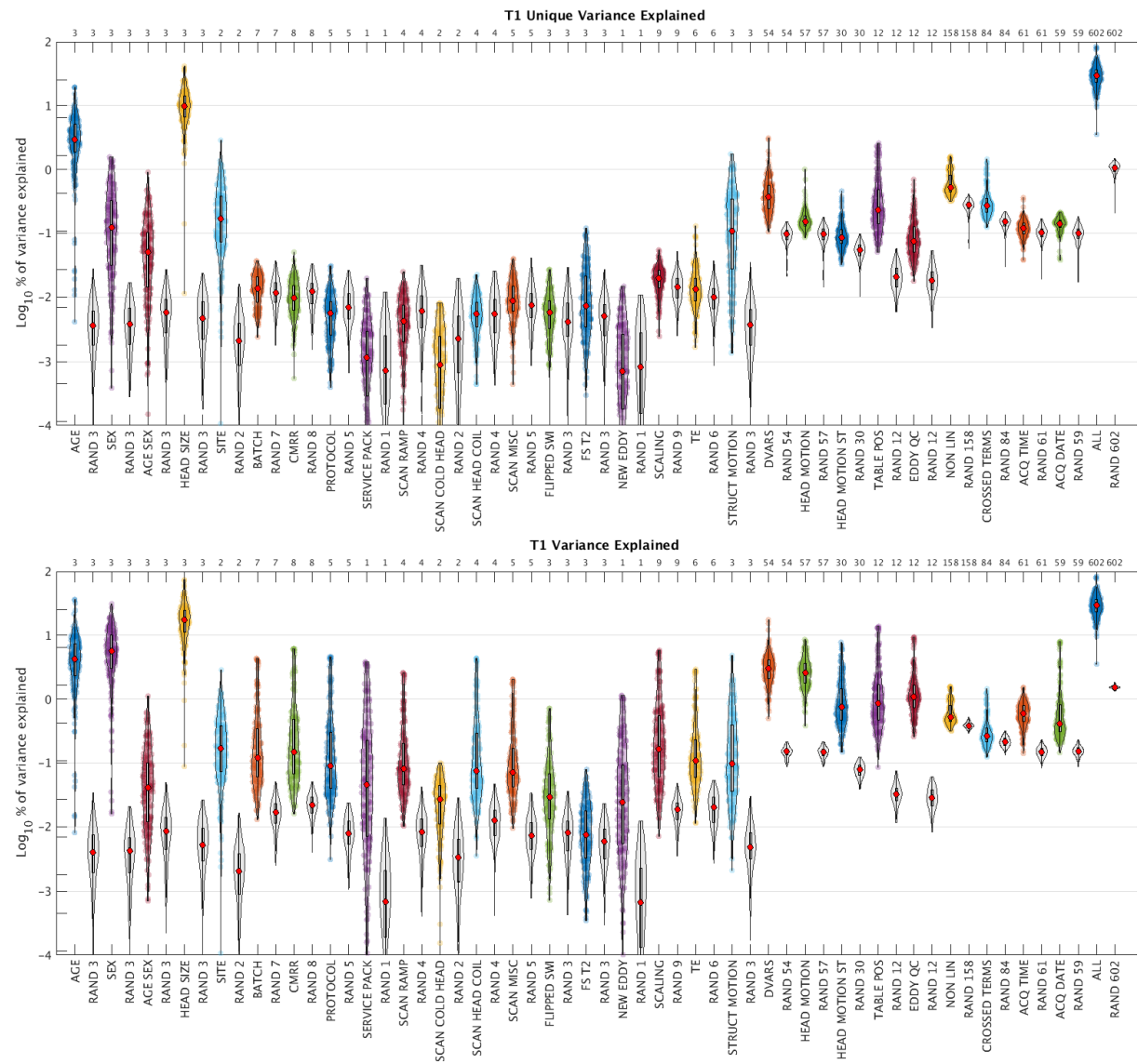


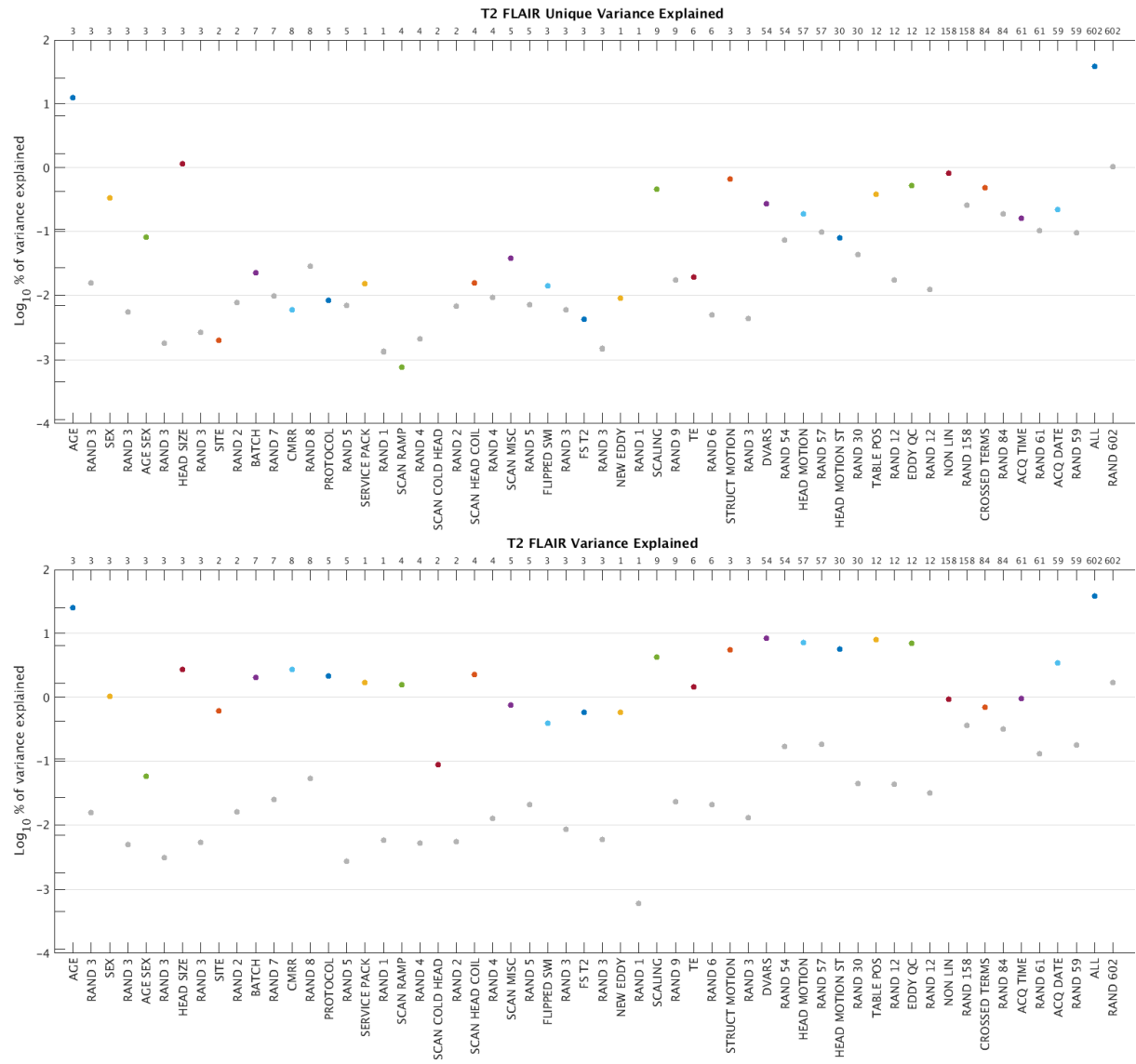


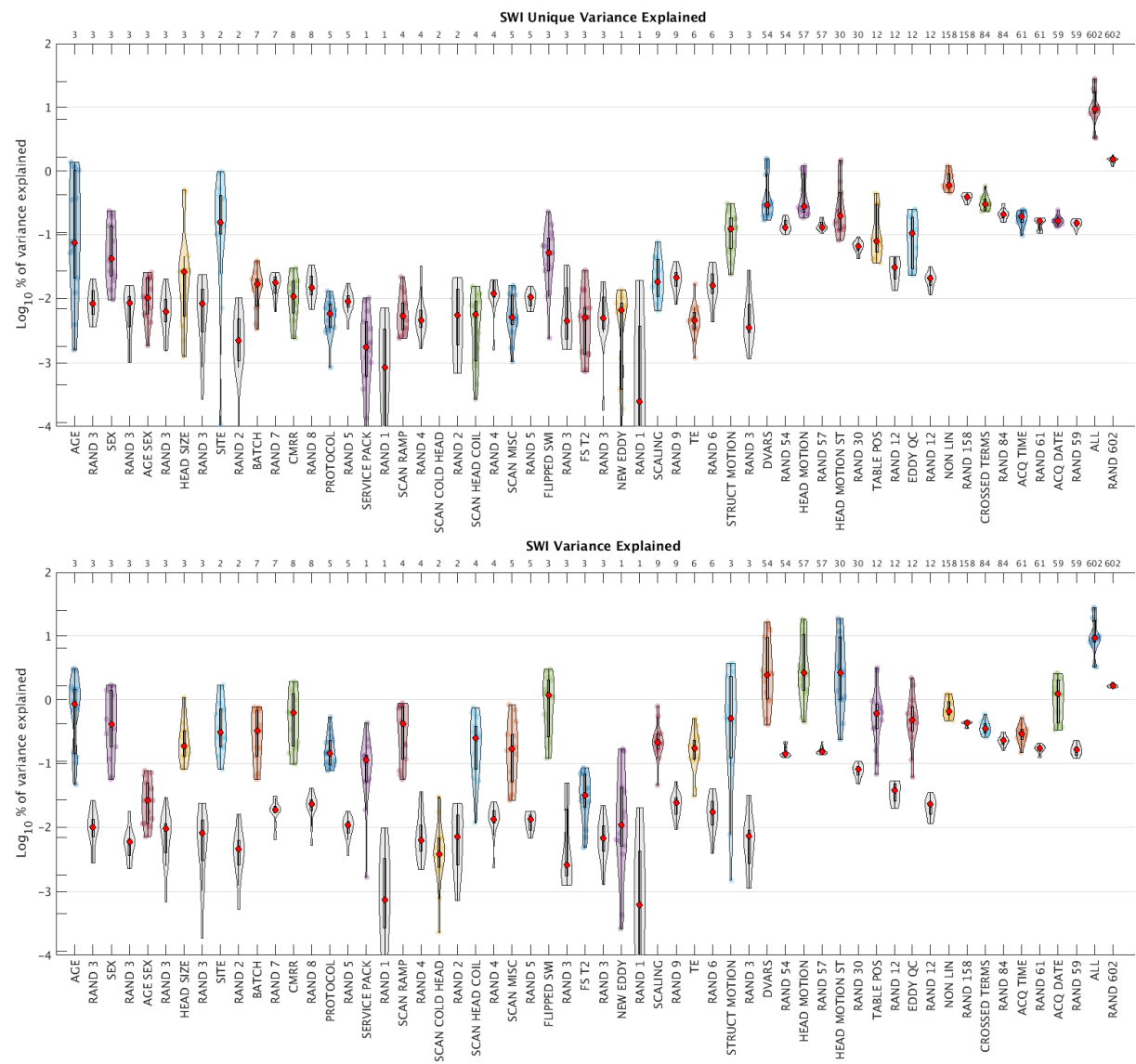
A.7 Violin plots for Variance Explained and Unique Variance Explained

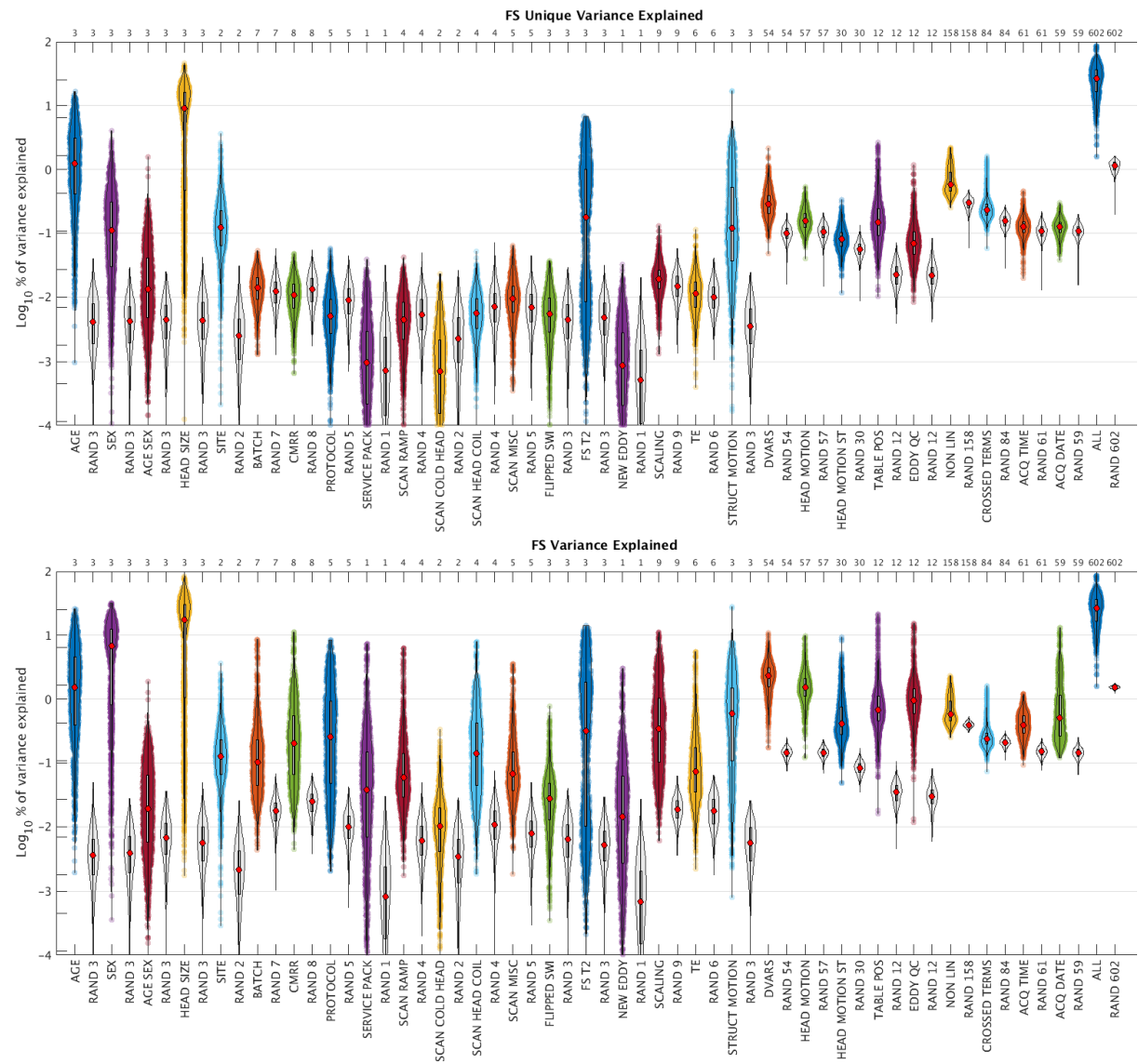


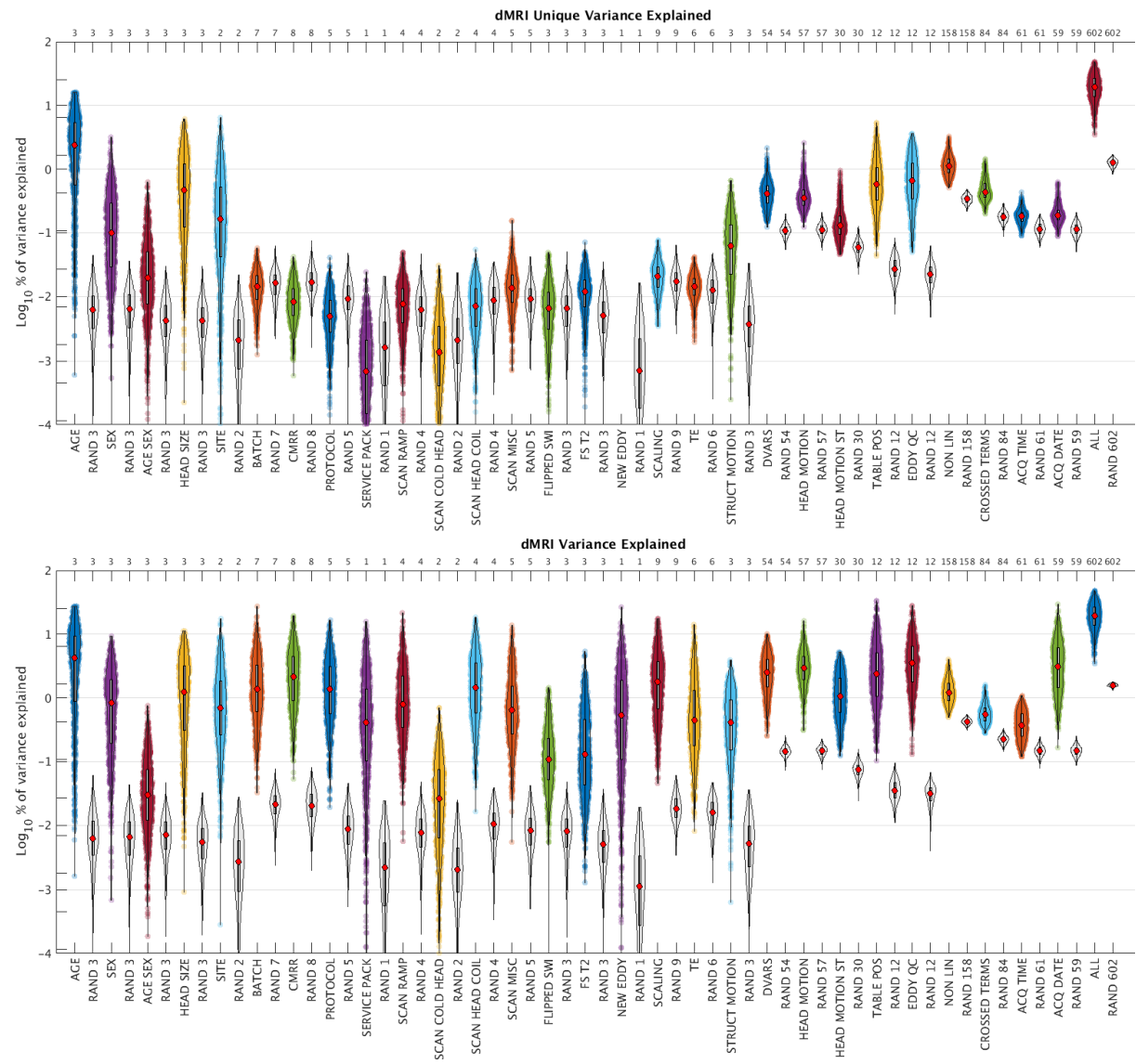


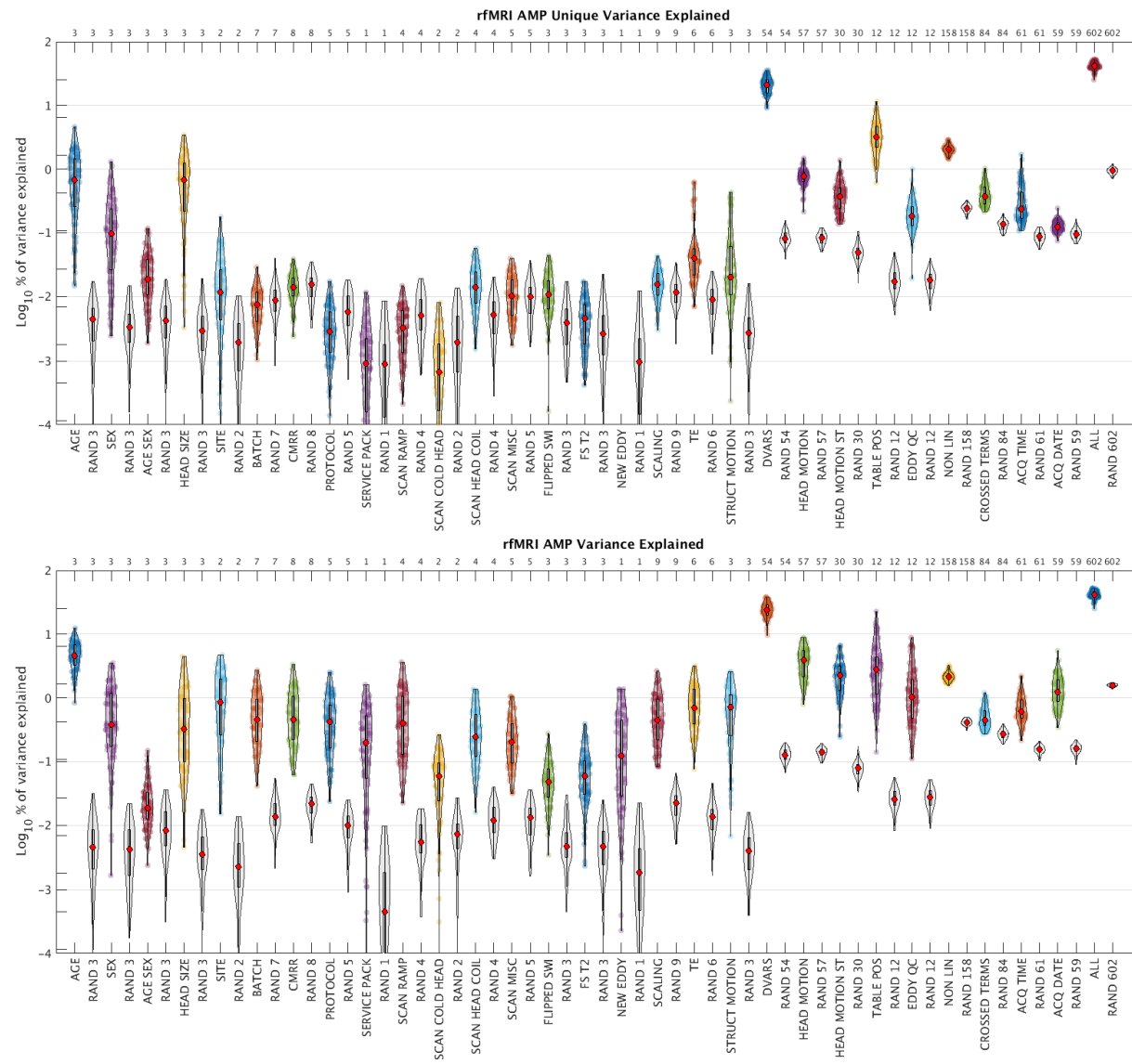


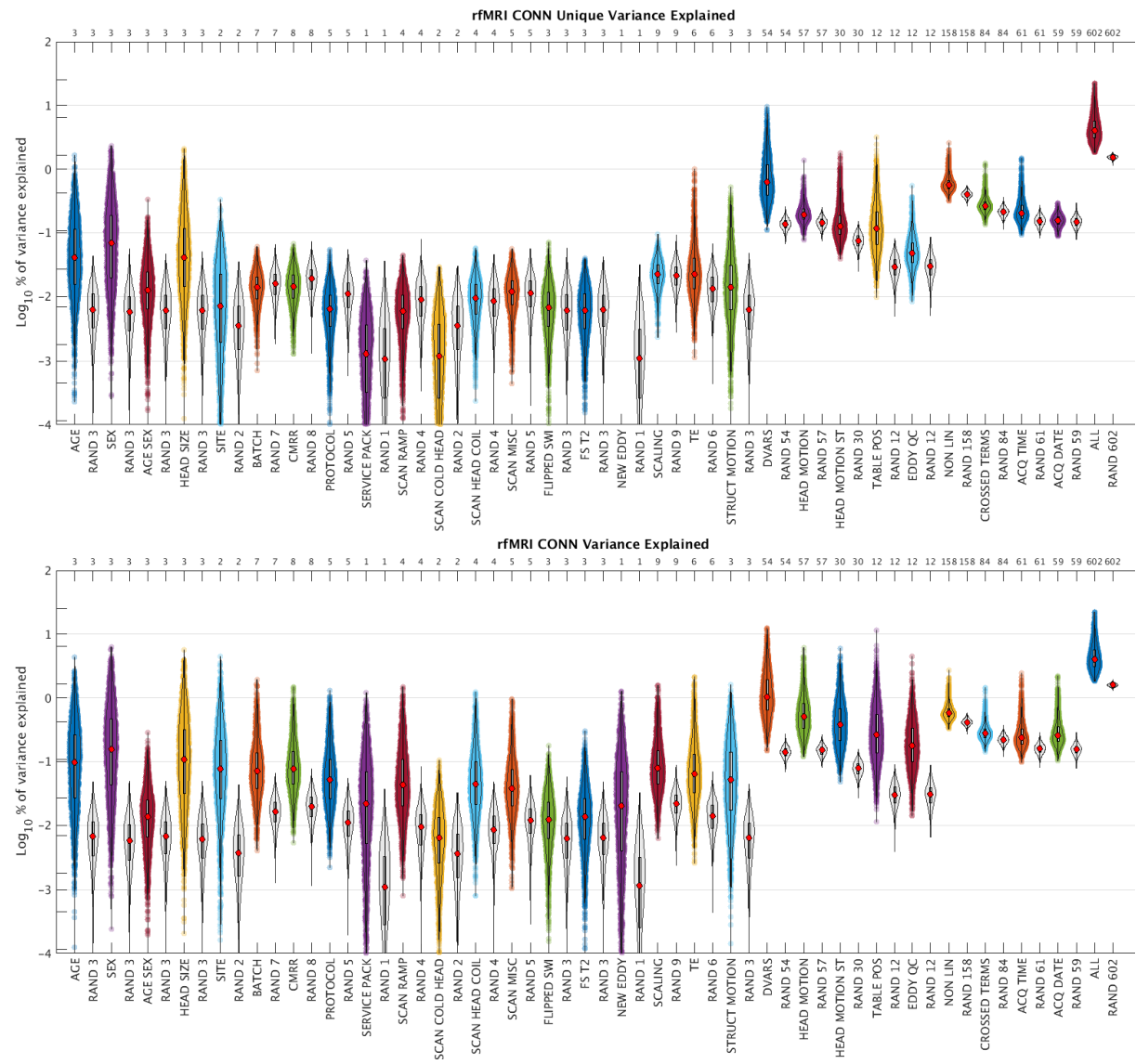


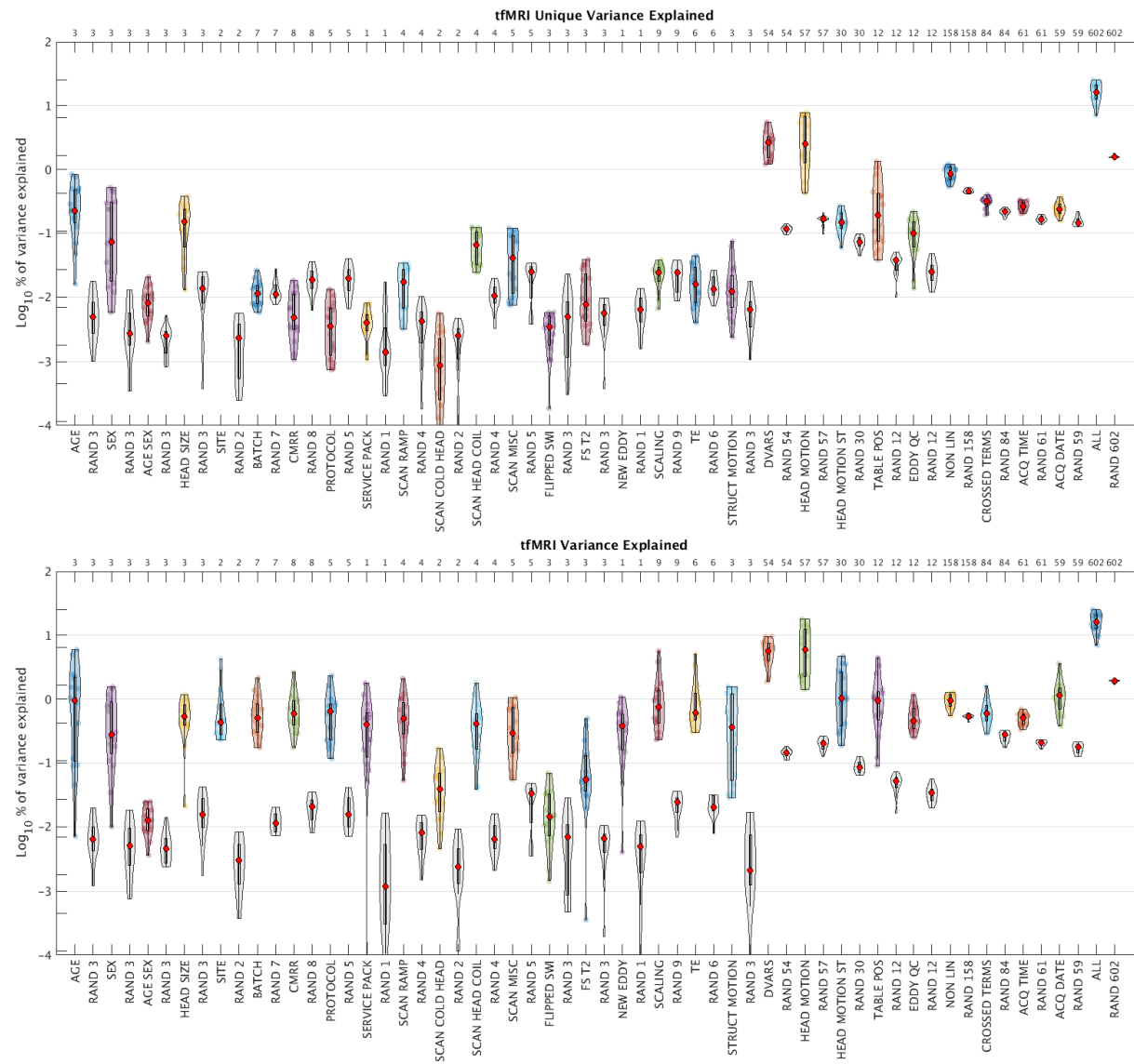




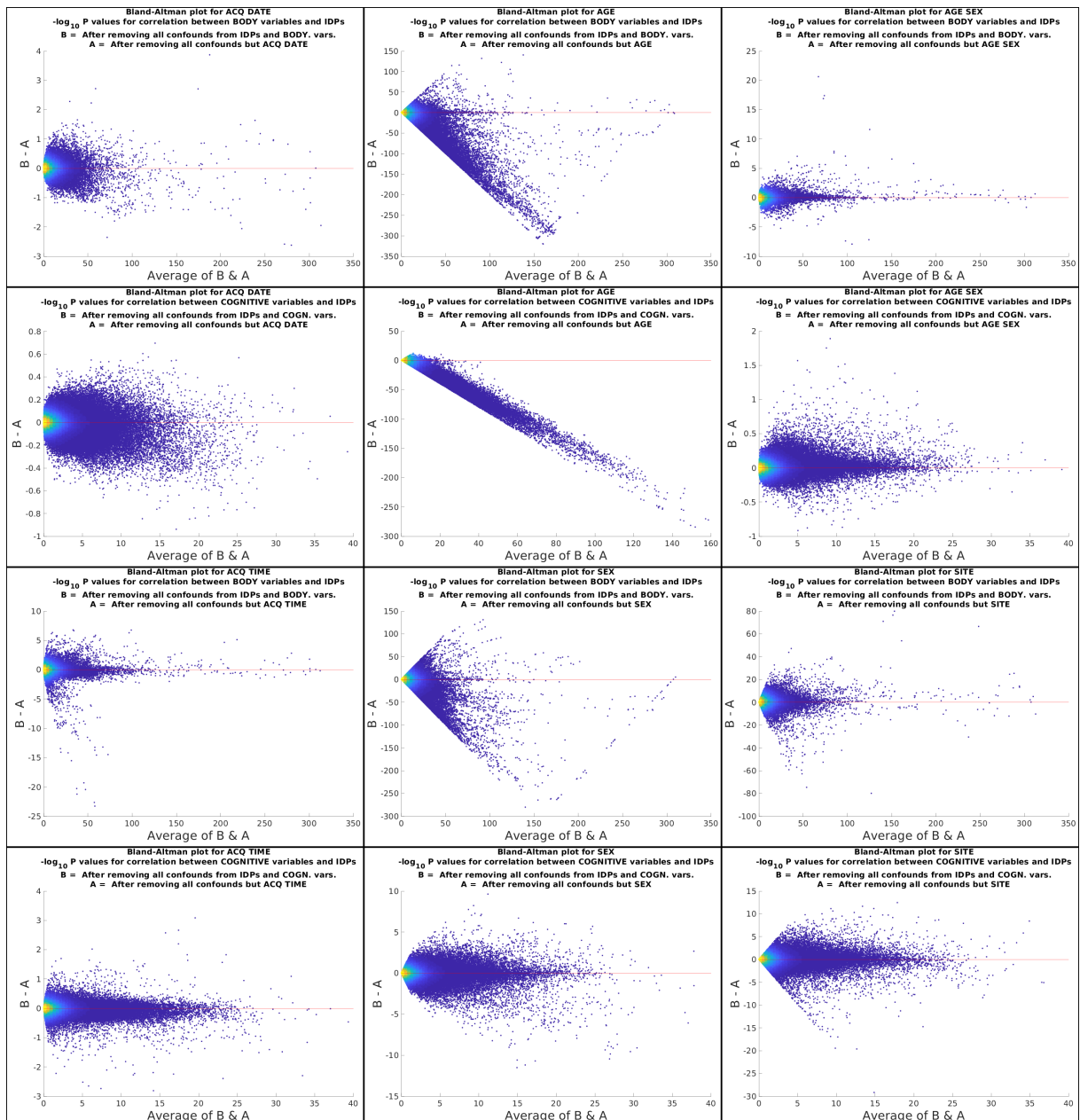


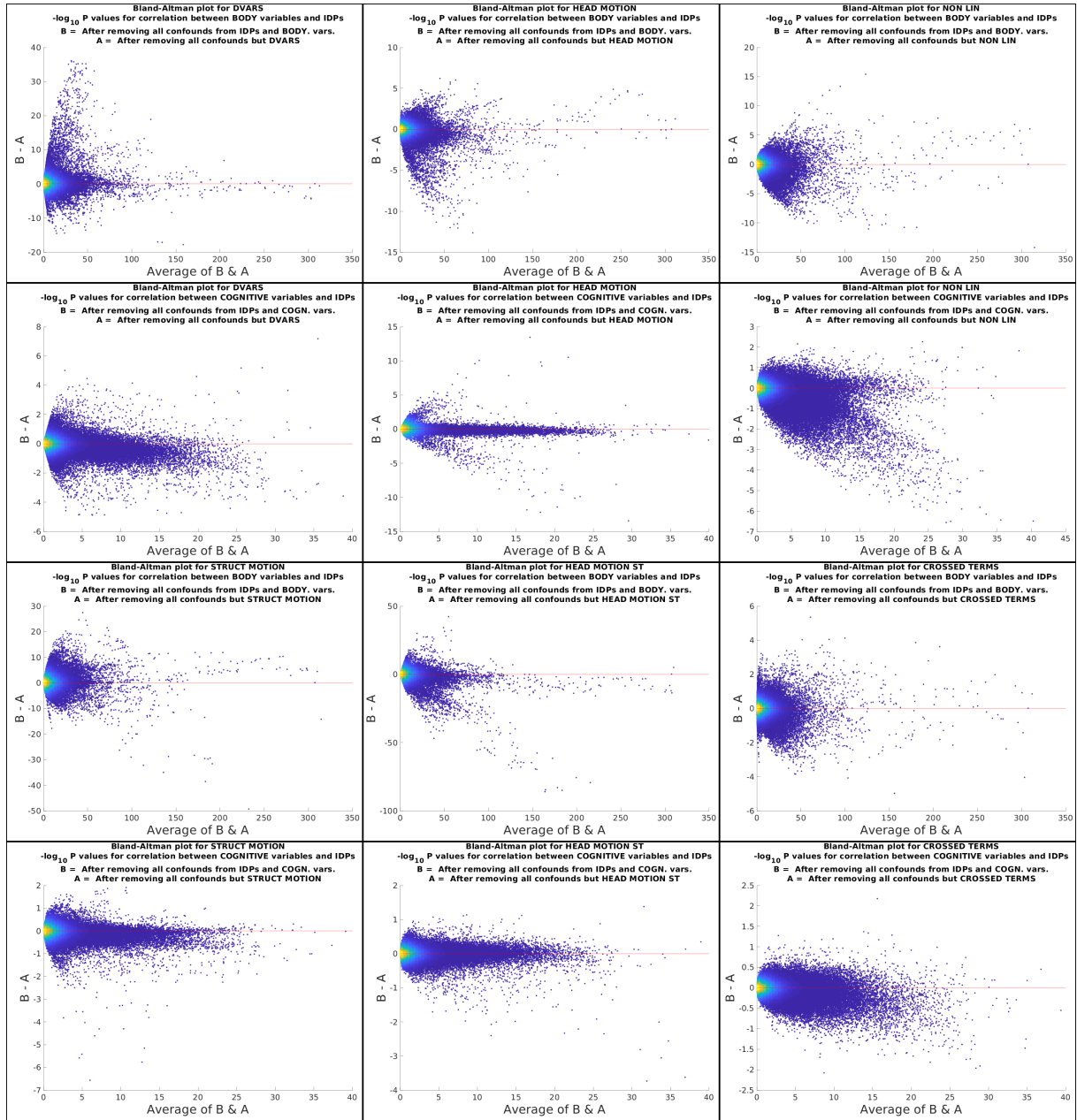


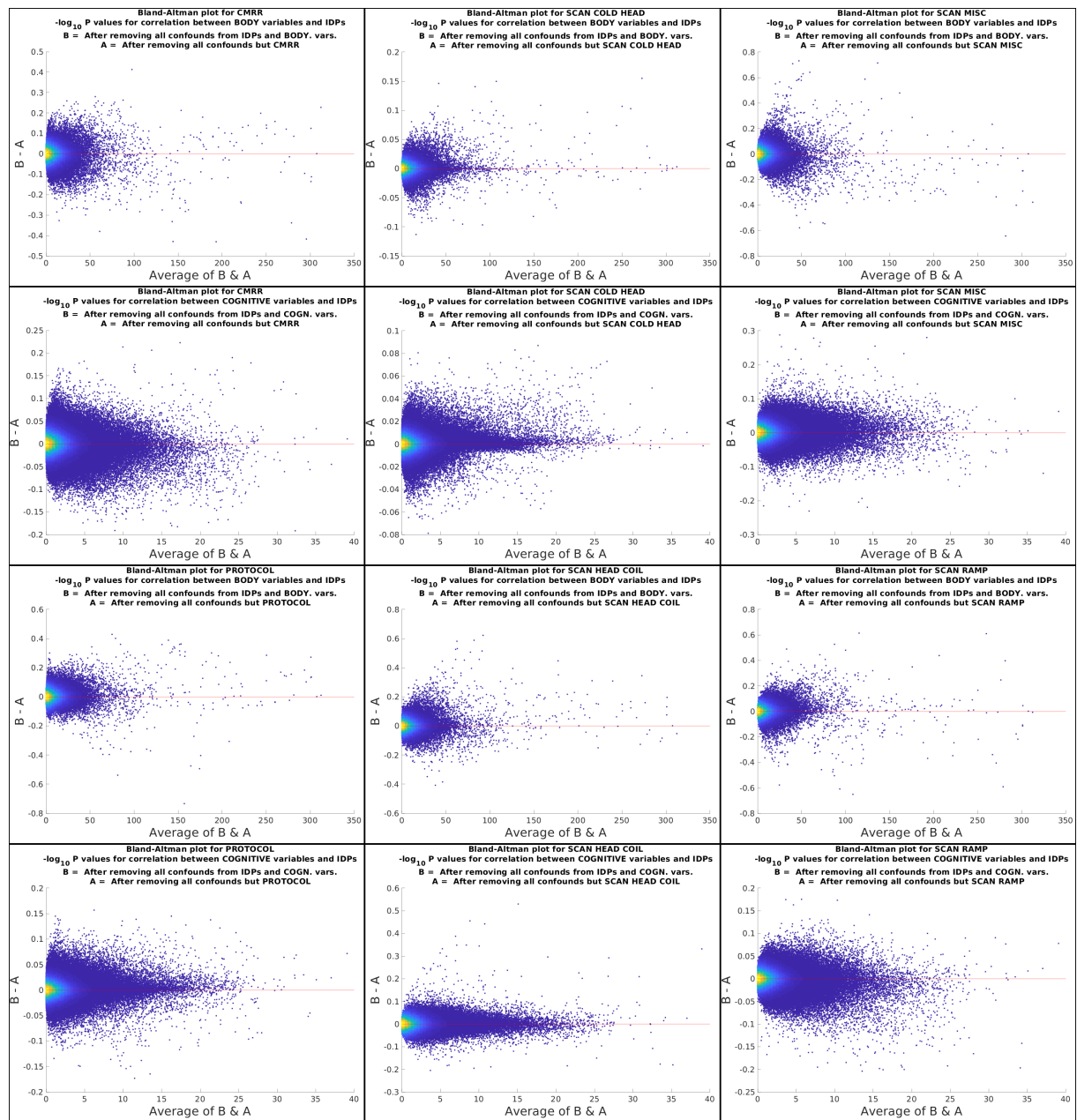


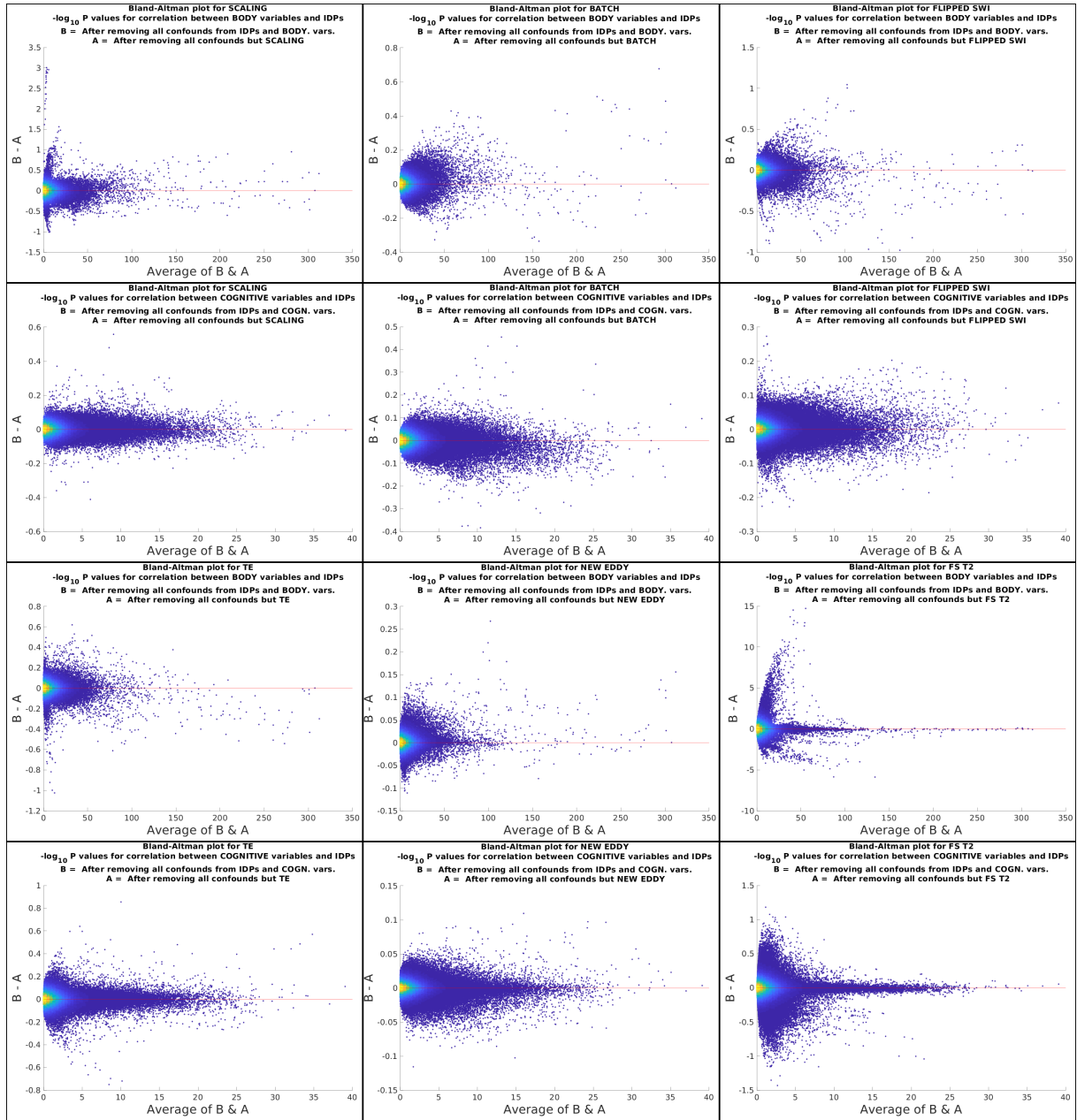


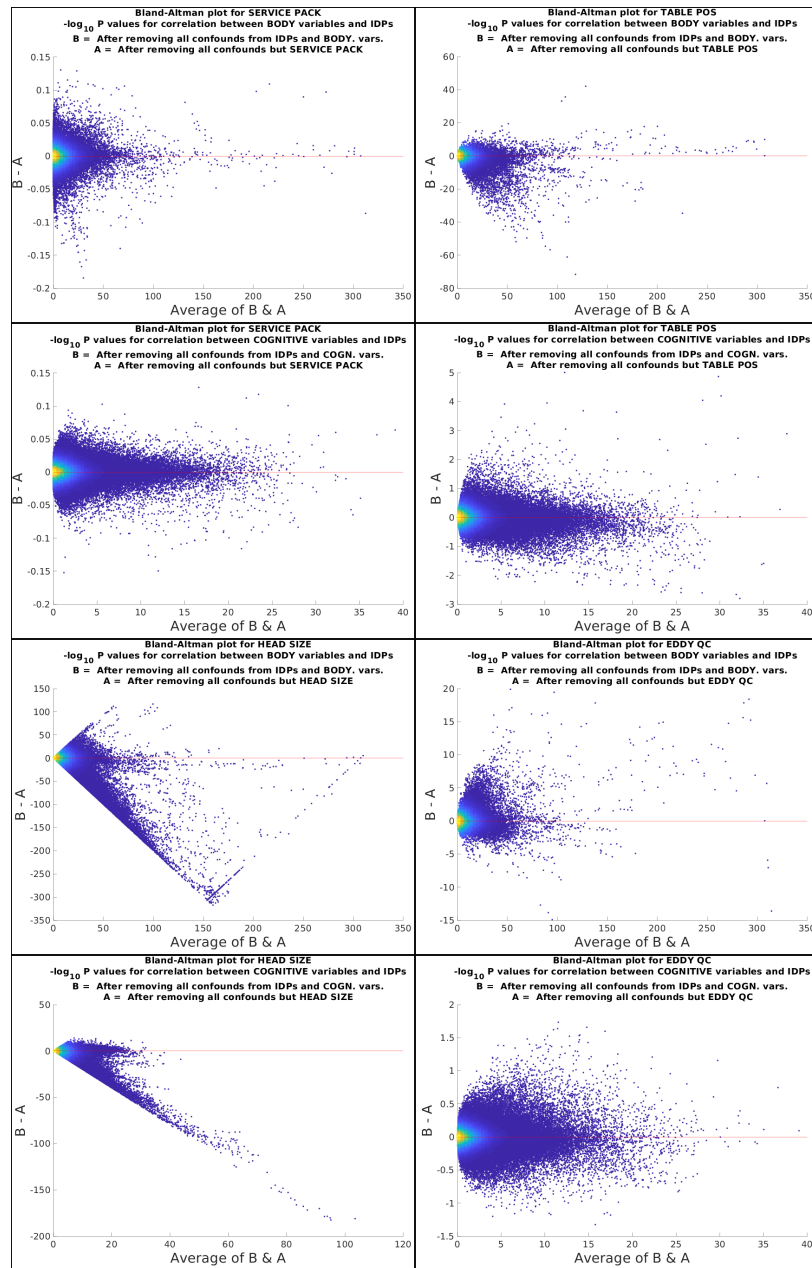
A.8 Bland-Altman plots



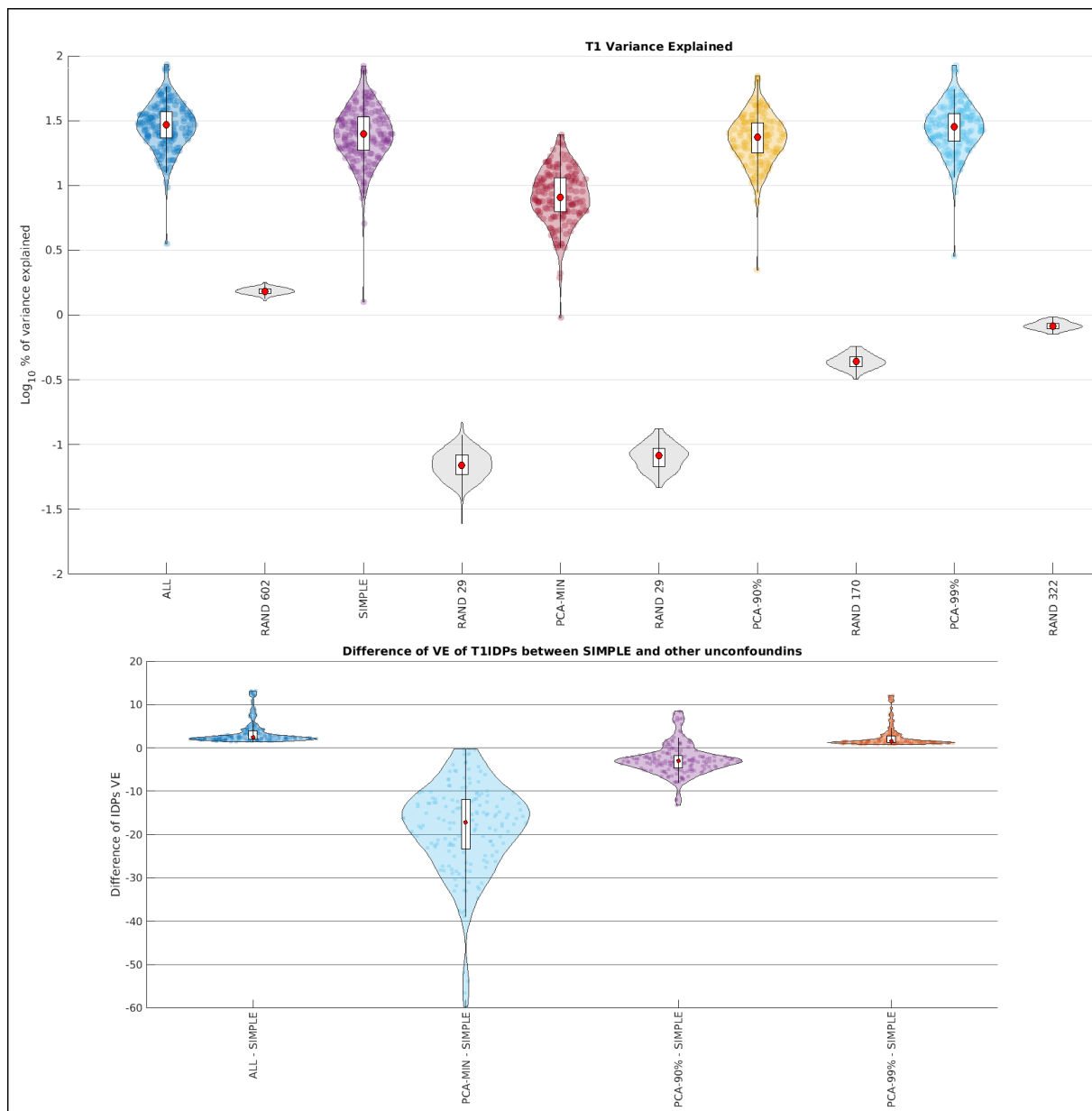


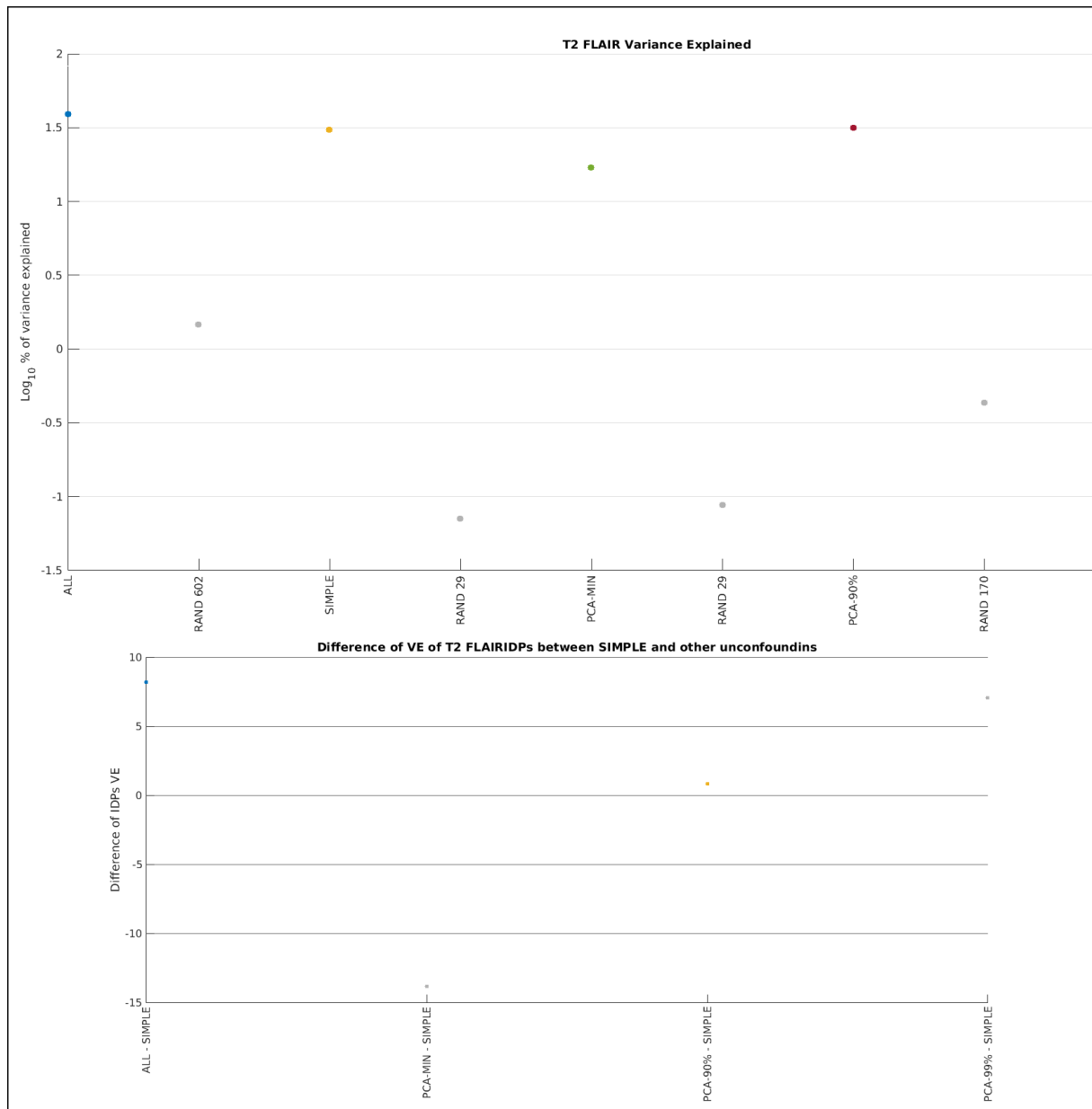


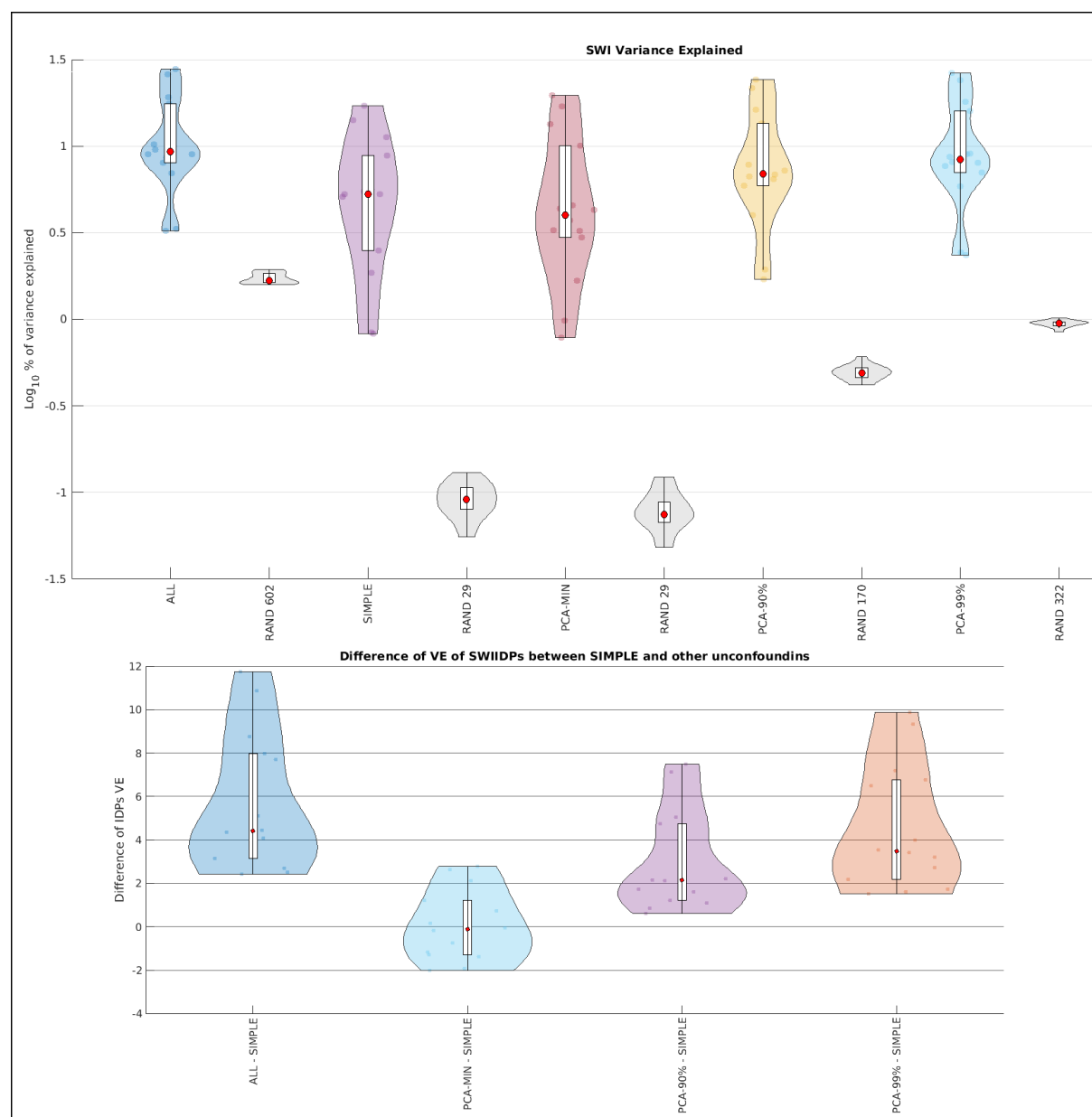


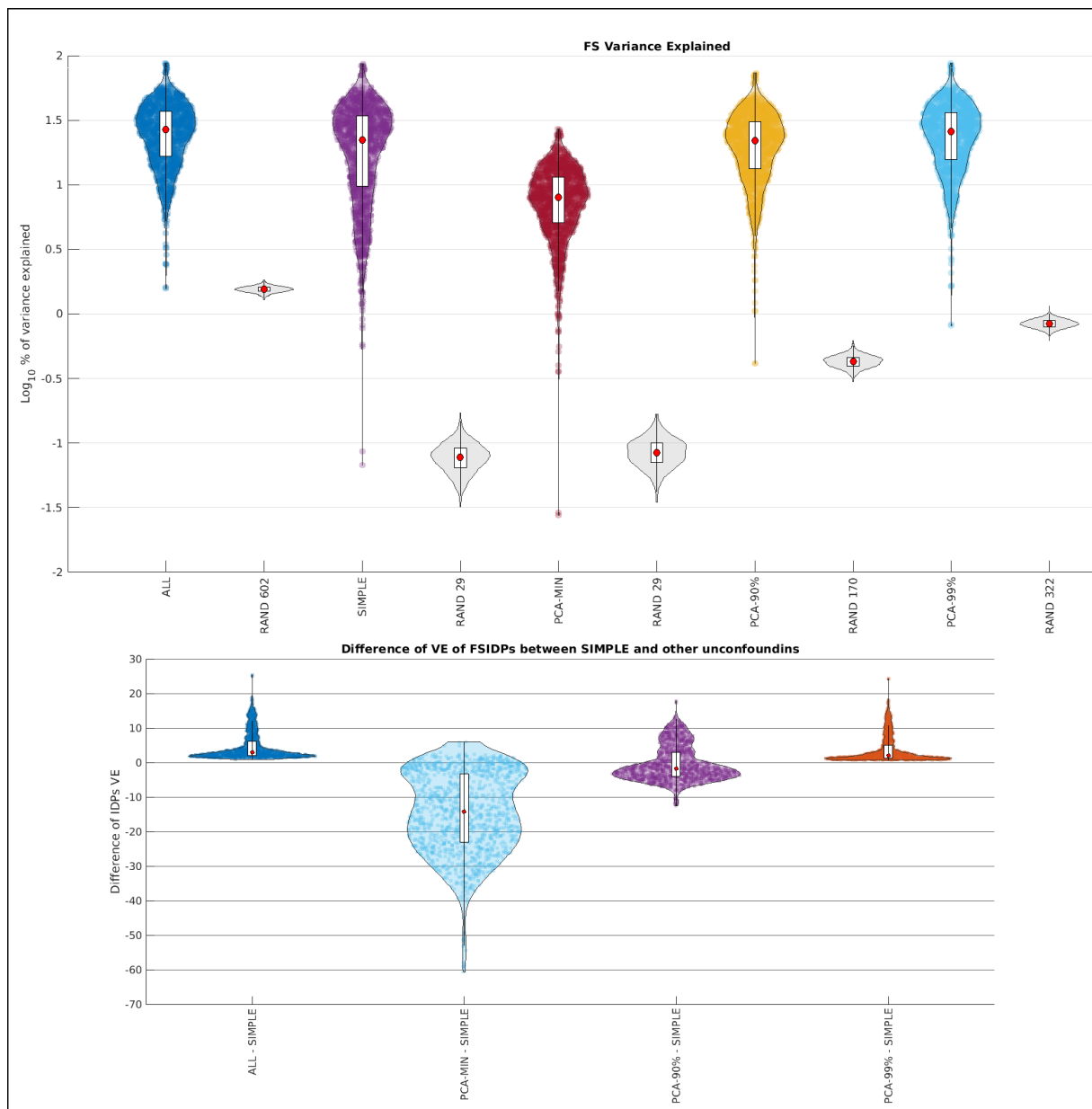


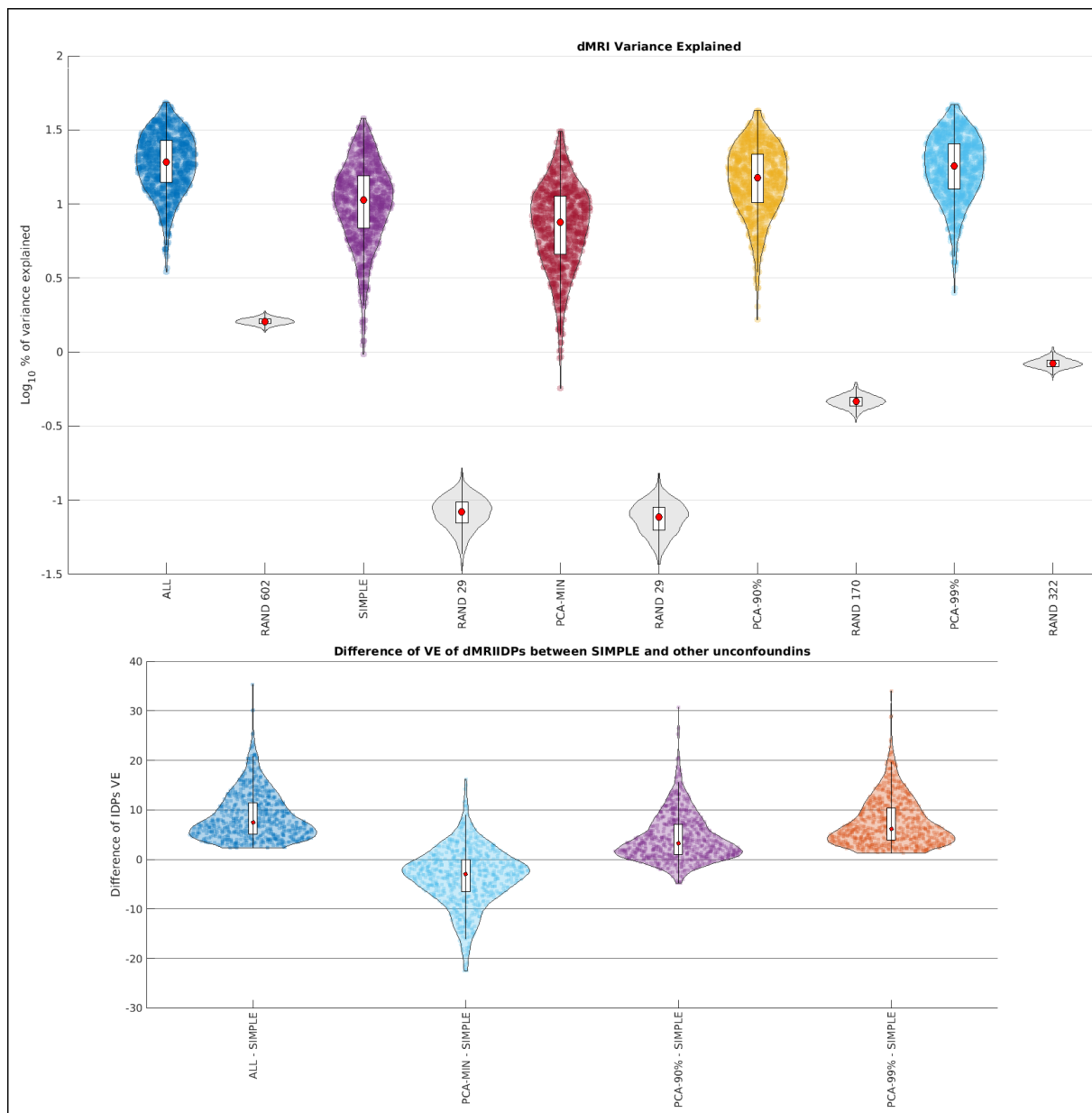
A.9 Variance of the IDPs explained by different unconfounding settings

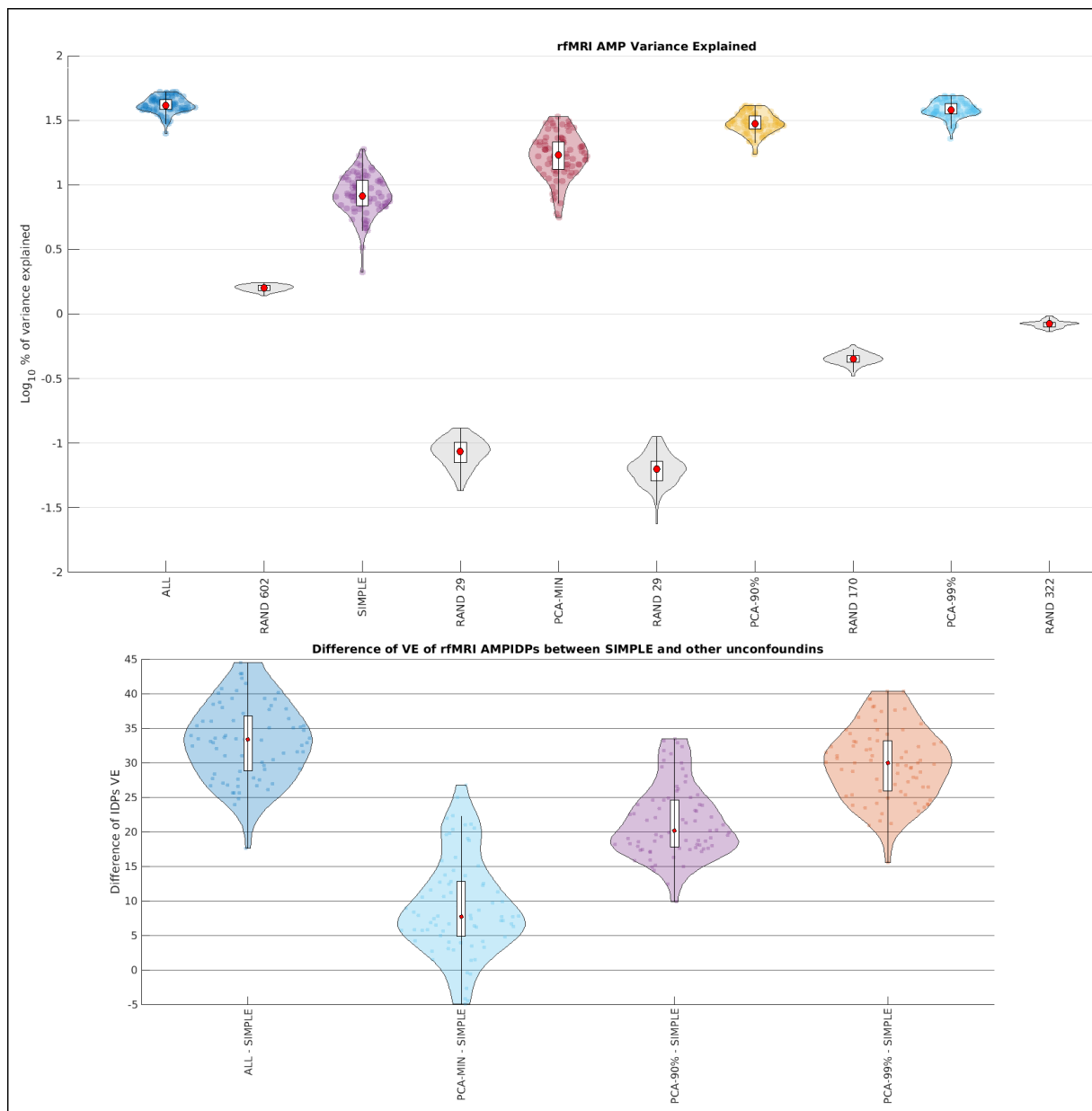


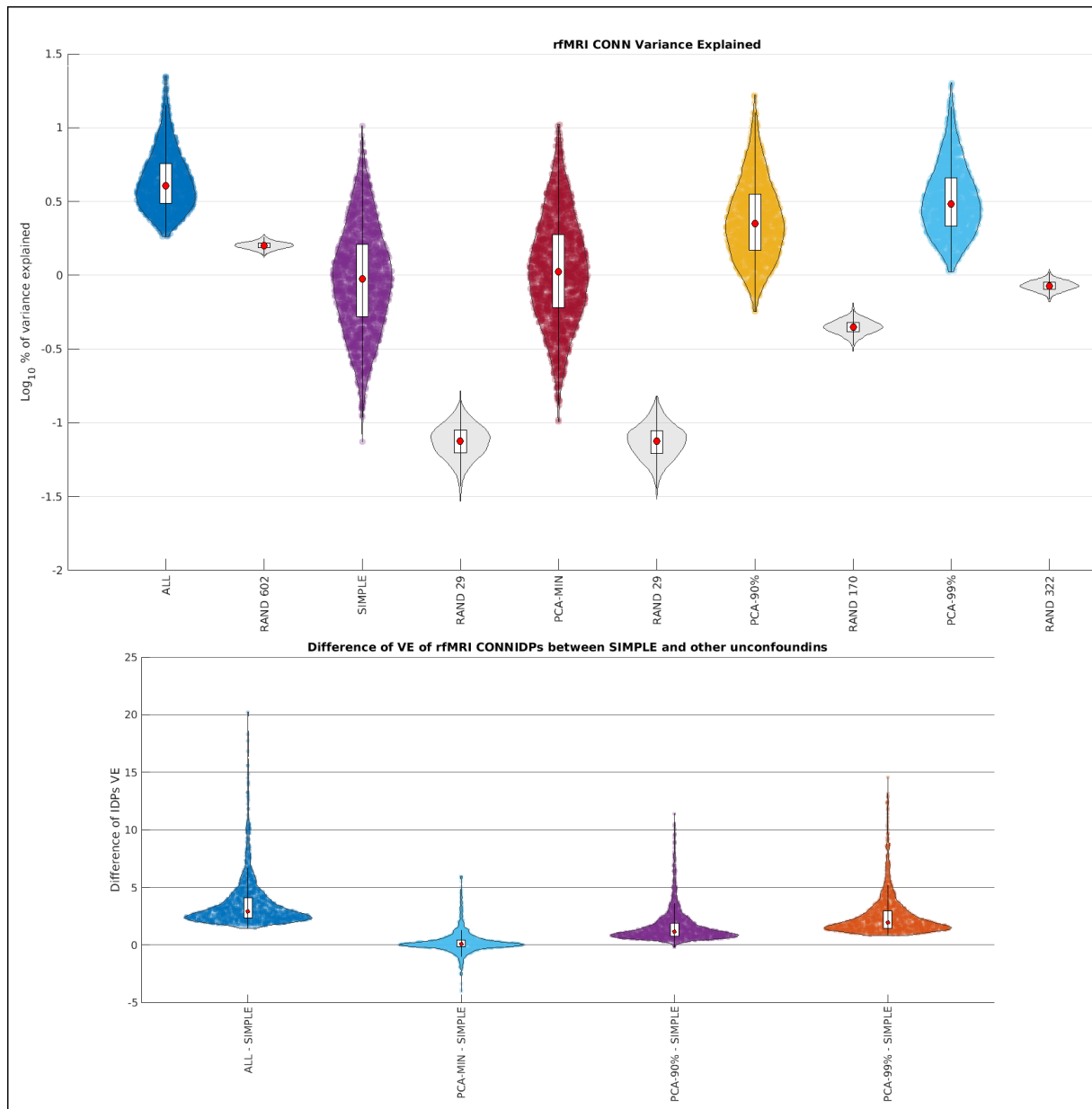


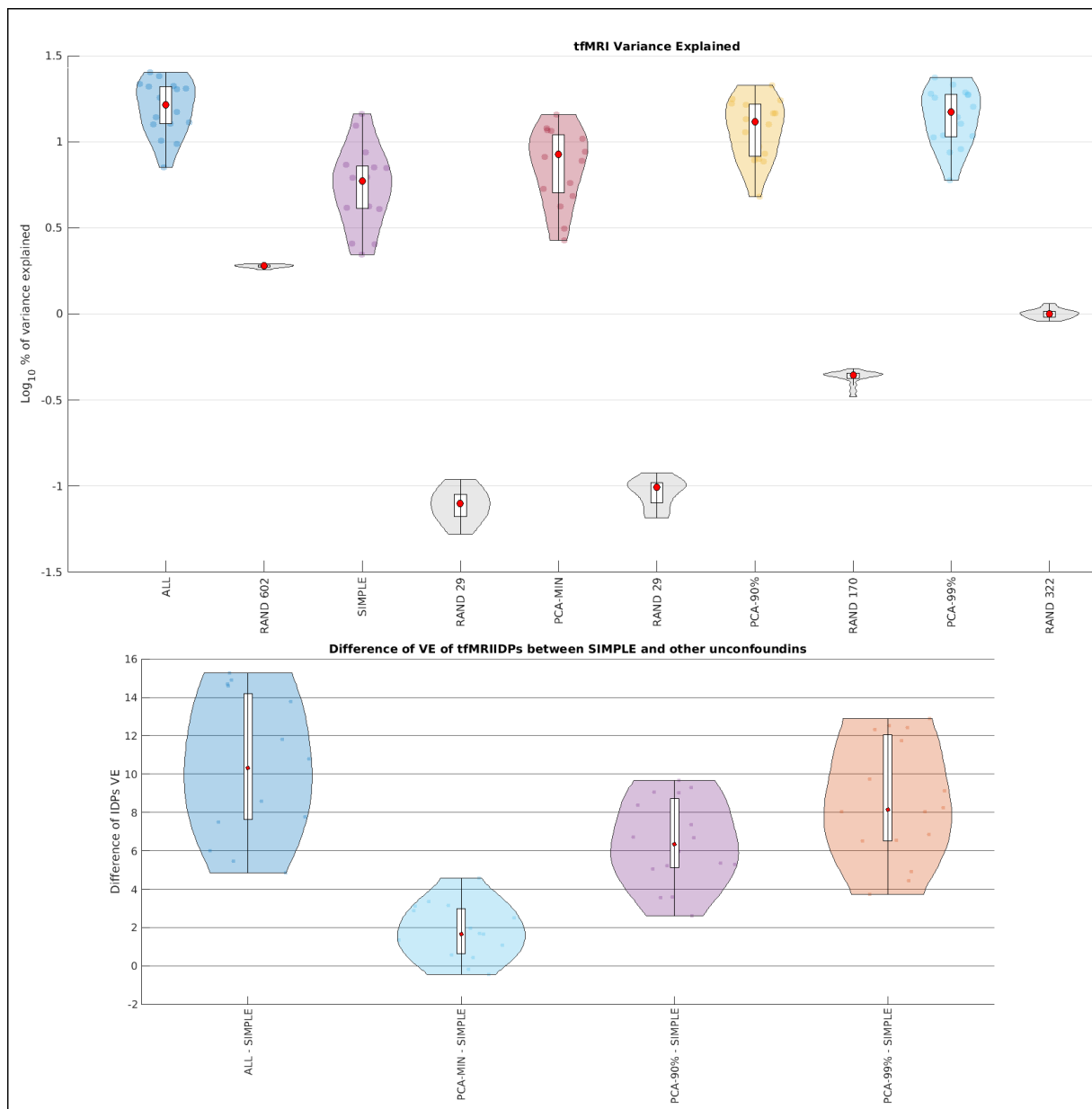




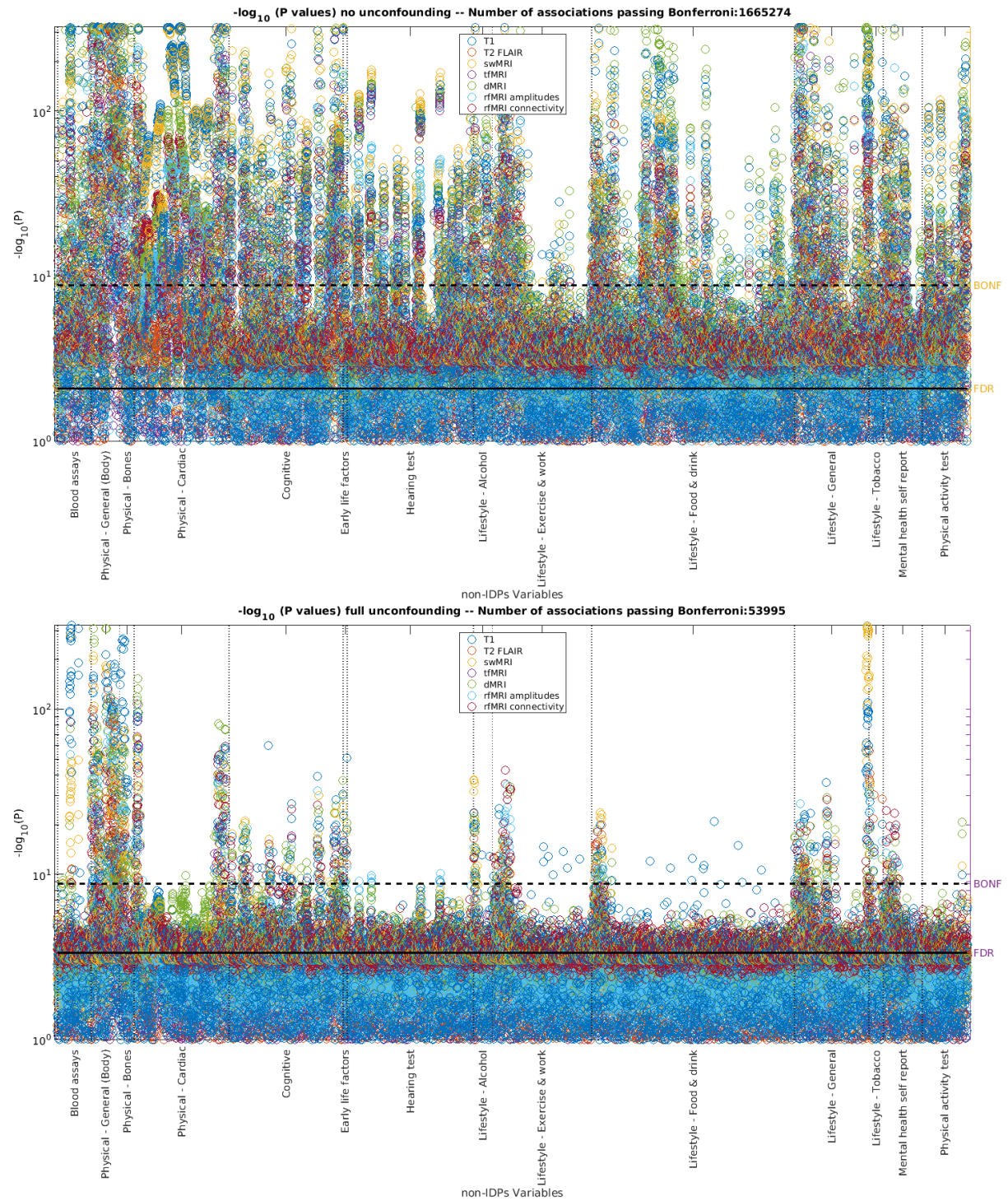


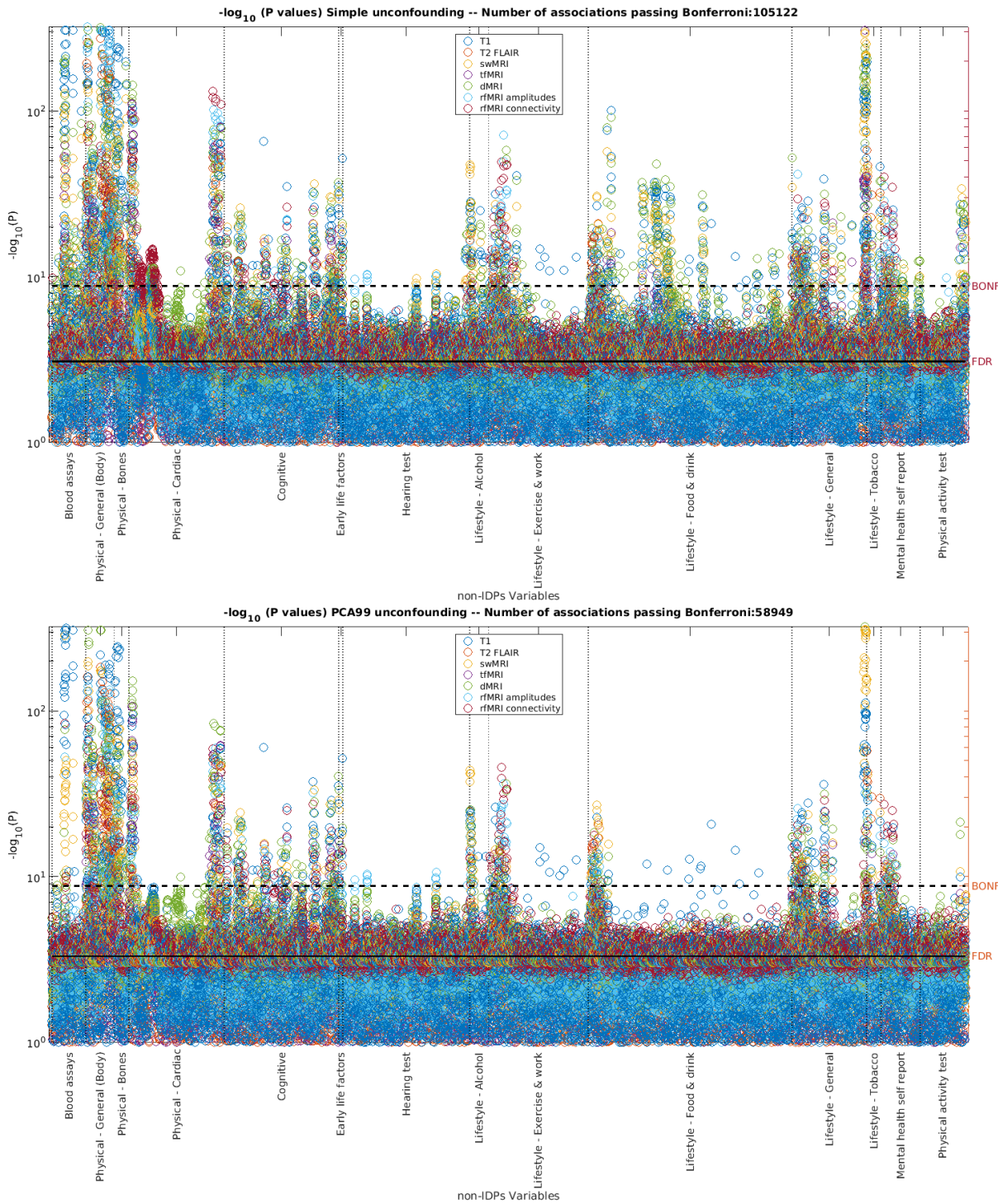


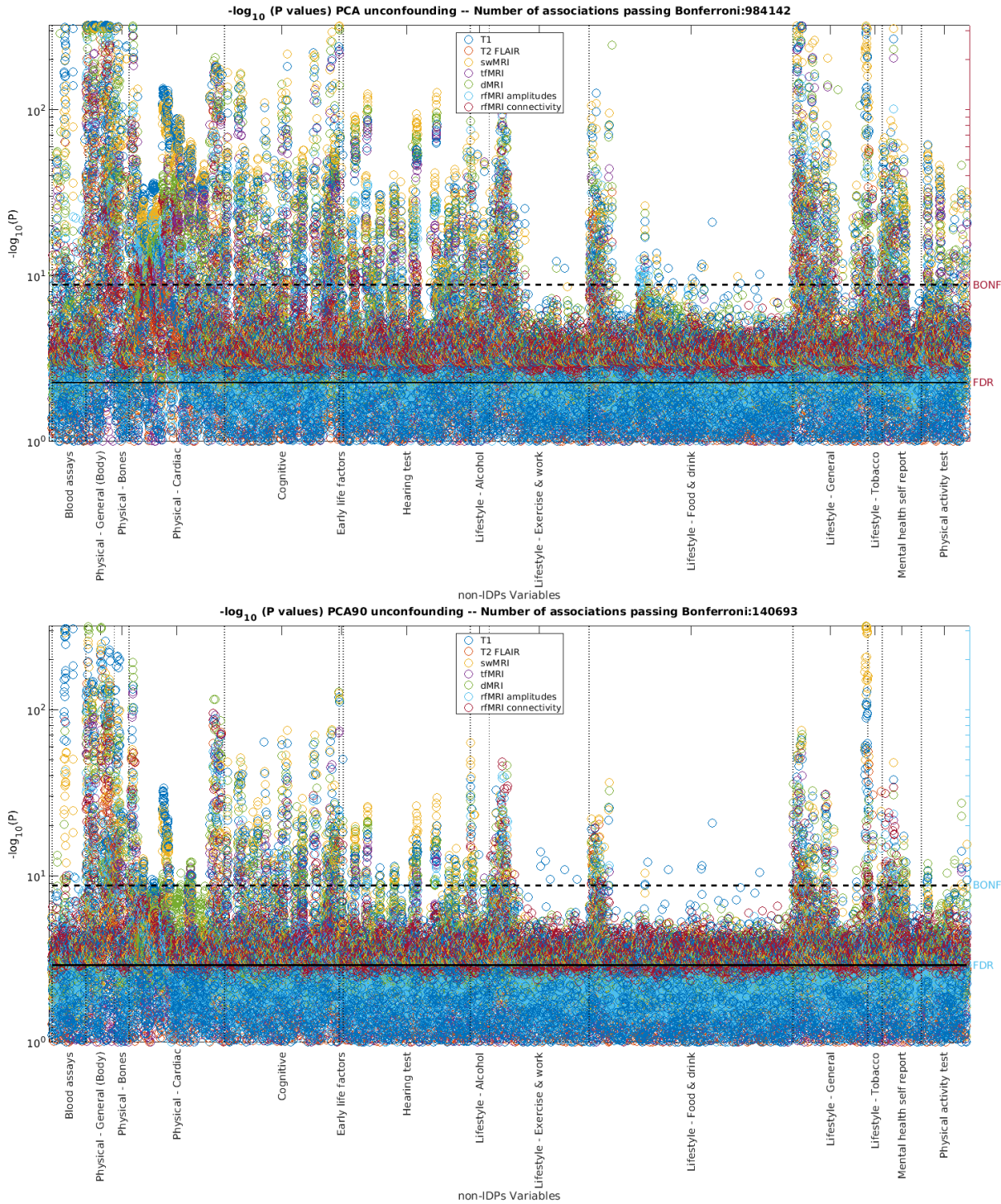




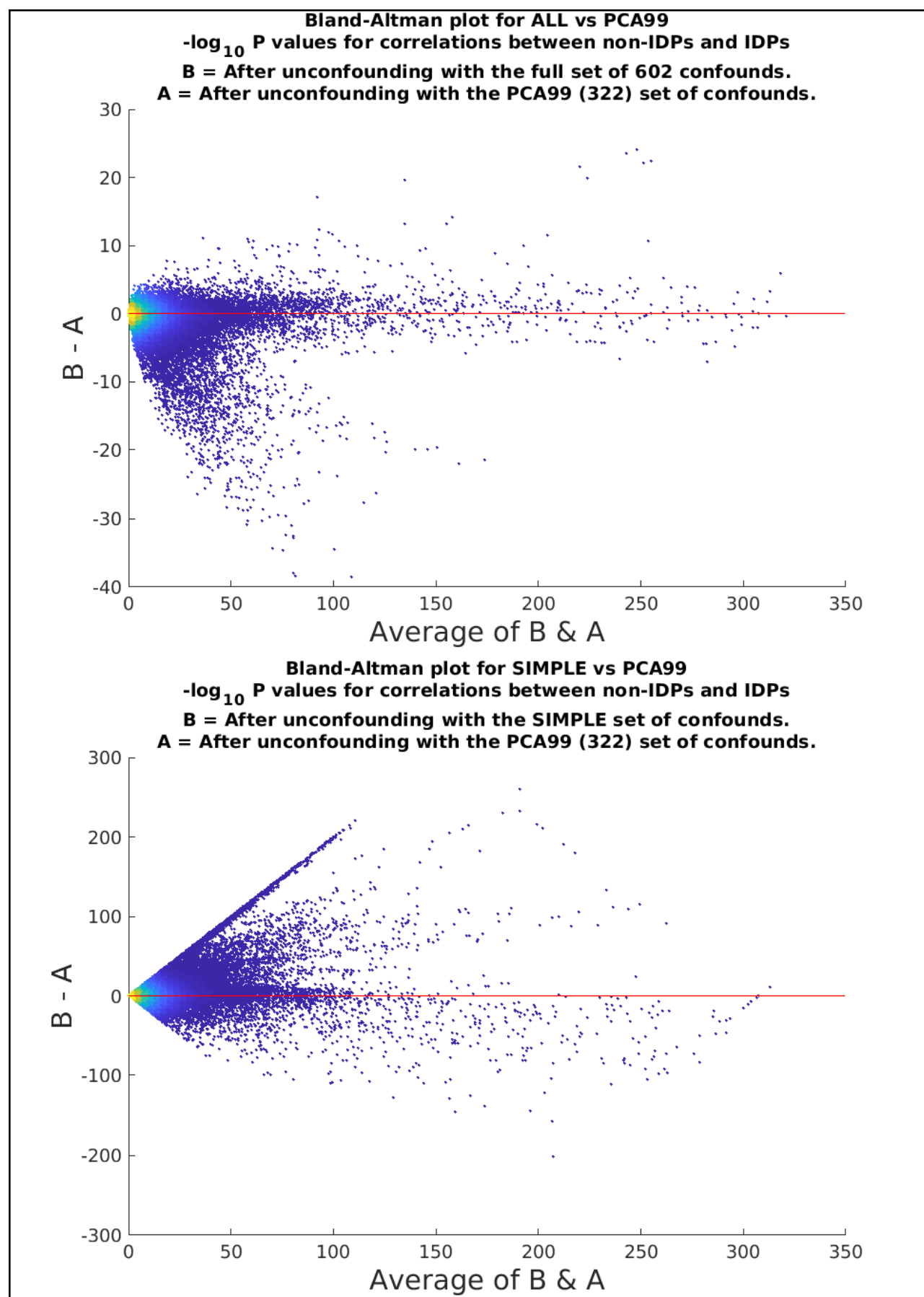
A.10 Manhattan plots of different unconfounding settings



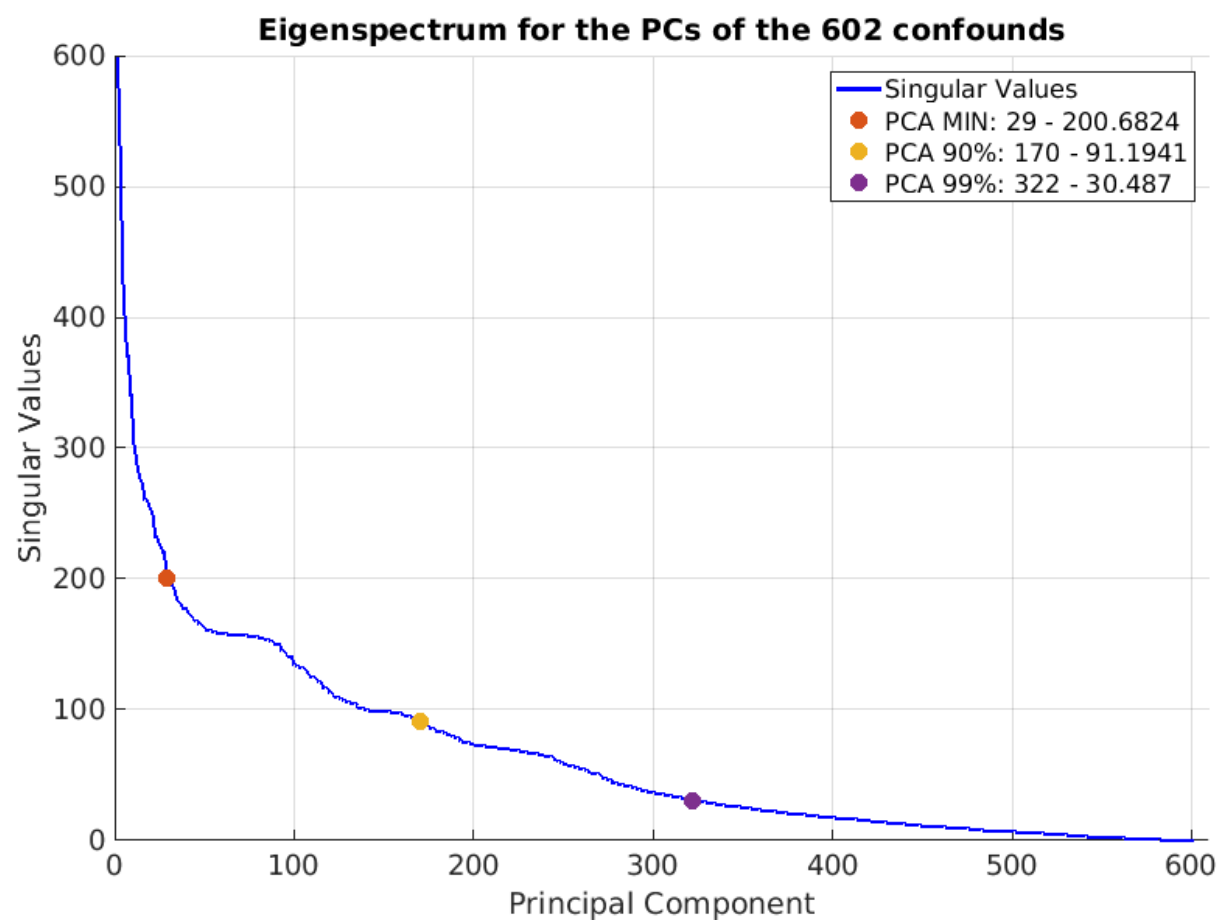




A.11 BA plots of unconfounding sets



A.12 Eigenspectrum for the Principal Components of the full confound set



Bibliography

- Abe, S., Irimia, A., and Van Horn, J. D. (2015). Quality Control Considerations for the Effective Integration of Neuroimaging Data. In *Data Integration in the Life Sciences*, pages 195–201. Springer.
- Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., and Calhoun, V. (2020). Hype versus hope: Deep learning encodes more predictive and robust brain imaging representations than standard machine learning. preprint, Neuroscience.
- Afyouni, S. and Nichols, T. E. (2018). Insight and inference for DVARS. *Neuroimage*, 172:291–312.
- Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., Miller, K. L., and Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, 166:400–424.
- Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M., Miller, K. L., Nichols, T. E., and Smith, S. M. (2020). Confound modelling in uk biobank brain imaging. *NeuroImage*, page 117002.
- Alloza, C., Blesa-Cábez, M., Bastin, M. E., Madole, J. W., Buchanan, C. R., Janssen, J., Gibson, J., Deary, I. J., Tucker-Drob, E. M., Whalley, H. C., Arango, C., McIntosh, A. M., Cox, S. R., and Lawrie, S. M. (2020). Psychotic-like experiences, polygenic risk scores for schizophrenia, and structural properties of the salience, default mode, and central-executive networks in healthy participants from UK Biobank. *Translational Psychiatry*, 10(1):122.
- Alnæs, D., Kaufmann, T., van der Meer, D., Córdova-Palomera, A., Rokicki, J., Moberget, T., Bettella, F., Agartz, I., Barch, D. M., Bertolino, A., et al. (2019). Brain heterogeneity in schizophrenia and its association with polygenic risk. *JAMA psychiatry*, 76(7):739–748.
- Ambikairajah, A., Tabatabaei-Jafari, H., Walsh, E., Hornberger, M., and Cherbuin, N. (2020). Longitudinal Changes in Fat Mass and the Hippocampus. *Obesity*, 28(7):1263–1269.
- Anatürk, M., Suri, S., Smith, S. M., Ebmeier, K. P., and Sexton, C. E. (2020). Mid-life and late life activities and their relationship with MRI measures of brain structure and functional connectivity in the UK Biobank cohort. preprint, Neuroscience.
- Anderson, K. M., Collins, M. A., Chin, R., Ge, T., Rosenberg, M. D., and Holmes, A. J. (2018). The transcriptional landscape of cortical interneurons underlies *in-vivo* brain function and schizophrenia risk. preprint, Neuroscience.
- Anderson, K. M., Collins, M. A., Chin, R., Ge, T., Rosenberg, M. D., and Holmes, A. J. (2020). Transcriptional and imaging-genetic association of cortical interneurons, brain function, and schizophrenia risk. *Nature Communications*, 11(1):2889.
- Andersson, J. L., Graham, M. S., Zsoldos, E., and 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., and Smith, S. (2007a). Non-linear registration aka Spatial normalisation. Internal Technical Report TR07JA2, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK. Available at www.fmrib.ox.ac.uk/analysis/techrep for downloading.
- Andersson, J. L. and Skare, S. (2010). Image distortion and its correction in diffusion MRI. In Jones, D., editor, *Diffusion MRI: theory, methods, and applications*, pages 285–302. Oxford University Press, Oxford.
- Andersson, J. L., Skare, S., and 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., Smith, S., and Jenkinson, M. (2007b). Non-Linear Optimisation. Internal Technical Report TR07JA1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK. Available at www.fmrib.ox.ac.uk/analysis/techrep for downloading.
- Andersson, J. L. and Sotiropoulos, S. N. (2015). Non-parametric representation and prediction of single- and multi-shell diffusion-weighted MRI data using Gaussian processes. *NeuroImage*, 122:166–176.
- Andersson, J. L. and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125:1063–1078.
- Andersson, J. L. R., Graham, M. S., Drobniak, I., Zhang, H., Filippini, N., and Bastiani, M. (2017). Towards a comprehensive framework for movement and distortion correction of diffusion MR images: Within volume movement. *Neuroimage*, 152:450–466.

- Arslan, S., Ktena, S. I., Glocker, B., and Rueckert, D. (2018). Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity. In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, pages 3–13. Springer.
- Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.
- Baker, A. P., Brookes, M. J., Rezek, I. A., Smith, S. M., Behrens, T., Smith, P. J. P., and Woolrich, M. (2014). Fast transient networks in spontaneous human brain activity. *Elife*, 3:e01867.
- Barbu, M. C., Spiliopoulou, A., Colombo, M., McKeigue, P., Clarke, T.-K., Howard, D. M., Adams, M. J., Shen, X., Lawrie, S. M., McIntosh, A. M., and Whalley, H. C. (2020). Expression quantitative trait loci-derived scores and white matter microstructure in UK Biobank: a novel approach to integrating genetics and neuroimaging. *Translational Psychiatry*, 10(1):55.
- Barbu, M. C., Zeng, Y., Shen, X., Cox, S. R., Clarke, T.-K., Gibson, J., Adams, M. J., Johnstone, M., Haley, C. S., Lawrie, S. M., et al. (2019). Association of whole-genome and netrin1 signaling pathway-derived polygenic risk scores for major depressive disorder and white matter microstructure in the uk biobank. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(1):91–100.
- Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189.
- Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M., Lehmann, M., Hobbs, N., Clarkson, M. J., MacManus, D. G., Ourselin, S., and Fox, N. C. (2010). Head size, age and gender adjustment in MRI studies: a necessary nuisance? *Neuroimage*, 53(4):1244–1255.
- Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasrallah, I., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., et al. (2020). Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain*, 143(7):2312–2324.
- Basser, P. J., Mattiello, J., and LeBihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B*, 103(3):247–254.
- Bastiani, M., Cottaar, M., Fitzgibbon, S. P., Suri, S., Alfaro-Almagro, F., Sotiropoulos, S. N., Jbabdi, S., and Andersson, J. L. R. (2019). Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction. *Neuroimage*, 184:801–812.
- Bauermeister, S., Orton, C., Thompson, S., Barker, R. A., Bauermeister, J. R., Ben-Shlomo, Y., Brayne, C., Burn, D., Campbell, A., Calvin, C., et al. (2020). The dementias platform uk (dpuk) data portal. Technical Report 6, Neuroscience.
- Bearden, C. E. and Thompson, P. M. (2017). Emerging global initiatives in neurogenetics: the enhancing neuroimaging genetics through meta-analysis (enigma) consortium. *Neuron*, 94(2):232–236.
- Beaudet, G., Tsuchida, A., Petit, L., Tzourio, C., Caspers, S., Schreiber, J., Pausova, Z., Patel, Y., Paus, T., Schmidt, R., et al. (2020). Age-related changes of peak width skeletonized mean diffusivity (psmd) across the adult lifespan: A multi-cohort study. *Frontiers in Psychiatry*, 11:342.
- Beckmann, C. and Smith, S. (2004). Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Trans. on Medical Imaging*, 23(2):137–152.
- Behrens, T., Berg, H. J., Jbabdi, S., Rushworth, M., and Woolrich, M. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *Neuroimage*, 34(1):144–155.
- Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., Matthews, P., Brady, J., and Smith, S. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic resonance in medicine*, 50(5):1077–1088.
- Bennett, C. M. and Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191(1):133–155.
- Bijsterbosch, J., Harrison, S., Duff, E., Alfaro-Almagro, F., Woolrich, M., and Smith, S. (2017). Investigations into within- and between-subject resting-state amplitude variations. *Neuroimage*, 159:57–69.
- Biobank (2014). Biobank protocol specifications. http://biobank.ctsu.ox.ac.uk/crystal/docs/bmri_v4_23092014.pdf. Accessed: 2017-02-01.
- Biobank (2016). Biobank First Public Imaging Release. <http://biobank.ctsu.ox.ac.uk/crystal/label.cgi>. Accessed: 2016-01-01.
- Biobank (2017). New data from brain imaging and on heart attacks and strokes available. <https://www.ukbiobank.ac.uk/2017/02/new-data-from-brain-imaging-and-on-heart-attacks-and-strokes-available/>. Accessed: 2016-02-13.
- Biton, A., Traut, N., Poline, J.-B., Aribisala, B. S., Bastin, M. E., Bülow, R., Cox, S. R., Deary, I. J., Fukunaga, M., Grabe, H. J., Hagenaars, S., Hashimoto, R., Kikuchi, M., Muñoz Maniega, S., Nauck, M., Royle, N. A., Teumer, A., Hernandez, M. V., Völker, U., Wardlaw, J. M., Wittfeld, K., Yamamori, H., Alzheimer’s Disease Neuroimaging Initiative, Bourgeron, T., and Toro, R. (2019). Polygenic architecture of human neuroanatomical diversity. preprint, Neuroscience.
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.
- Bouckaert, R. R. (2008). Bayesian Network Classifiers in Weka. <http://www.cs.waikato.ac.nz/>

- [~remco/weka.bn.pdf](#). Accessed: 2016-09-7.
- Bowring, A., Telschow, F., Schwartzman, A., and Nichols, T. E. (2019). Spatial confidence sets for raw effect size images. *NeuroImage*, 203:116187.
- Braga, R. M., Van Dijk, K. R. A., Polimeni, J. R., Eldaief, M. C., and Buckner, R. L. (2019). Parallel distributed networks resolved at high resolution reveal close juxtaposition of distinct regions. *Journal of Neurophysiology*, 121(4):1513–1534.
- Breteler, M. M., Stöcker, T., Pracht, E., Brenner, D., and Stirnberg, R. (2014). MRI in the Thineland study: A novel protocol for population neuroimaging. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 10(4):P92.
- Buchanan, C. R., Bastin, M. E., Ritchie, S. J., Liewald, D. C., Madole, J. W., Tucker-Drob, E. M., Deary, I. J., and Cox, S. R. (2020). The effect of network thresholding and weighting on structural brain networks in the UK Biobank. *NeuroImage*, 211:116443.
- Casey, B., 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.
- CBS (2016). Qualitative Quality Control Manual. http://cbs.fas.harvard.edu/usr/mcmains/CBS_MRI_Qualitative_Quality_Control_Manual.pdf. Accessed: 2016-04-22.
- Chappell, M. A., Groves, A. R., Whitcher, B., and Woolrich, M. W. (2008). Variational bayesian inference for a nonlinear forward model. *IEEE Transactions on Signal Processing*, 57(1):223–236.
- Chen, J., Liu, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., Gupta, C. N., Franke, B., and Turner, J. A. (2014). Exploration of scanning effects in multi-site structural MRI studies. *Journal of neuroscience methods*, 230:37–50.
- Cheng, S., Wu, C., Qi, X., Liu, L., Ma, M., Zhang, L., Cheng, B., Liang, C., Li, P., Kafle, O. P., Wen, Y., and Zhang, F. (2020). A Large-Scale Genetic Correlation Scan Between Intelligence and Brain Imaging Phenotypes. *Cerebral Cortex*, 30(7):4197–4203.
- Cheng, W., Rolls, E. T., Ruan, H., and Feng, J. (2018). Functional Connectivities in the Brain That Mediate the Association Between Depressive Problems and Sleep Quality. *JAMA Psychiatry*, 75(10):1052.
- Chyzhyk, D., Varoquaux, G., Thirion, B., and Milham, M. (2018). Controlling a confound in predictive models with a test set minimizing its effect. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE.
- Cirulli, E. T., White, S., Read, R. W., Elhanan, G., Metcalf, W. J., Tanudjaja, F., Fath, D. M., Sandoval, E., Isaksson, M., Schlauch, K. A., Grzymalski, J. J., Lu, J. T., and Washington, N. L. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nature Communications*, 11(1):542.
- Cole, J. H. (2020). Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of Aging*, 92:34–42.
- Cook, P., Bai, Y., Nedjati-Gilani, S., Seunarine, K., Hall, M., Parker, G., and Alexander, D. (2006). Camino: open-source diffusion-MRI reconstruction and processing. In *14th scientific meeting of the international society for magnetic resonance in medicine*, volume 2759, page 2759. Seattle WA, USA.
- Cox, S., Ritchie, S., Fawns-Ritchie, C., Tucker-Drob, E., and Deary, I. (2019a). Structural brain imaging correlates of general intelligence in UK Biobank. *Intelligence*, 76:101376.
- Cox, S. R., Lyall, D. M., Ritchie, S. J., Bastin, M. E., Harris, M. A., Buchanan, C. R., Fawns-Ritchie, C., Barbu, M. C., de Nooij, L., Reus, L. M., Alloza, C., Shen, X., Neilson, E., Alderson, H. L., Hunter, S., Liewald, D. C., Whalley, H. C., McIntosh, A. M., Lawrie, S. M., Pell, J. P., Tucker-Drob, E. M., Wardlaw, J. M., Gale, C. R., and Deary, I. J. (2019b). Associations between vascular risk factors and brain MRI indices in UK Biobank. *European Heart Journal*, 40(28):2290–2300.
- Cox, S. R., Ritchie, S. J., Tucker-Drob, E. M., Liewald, D. C., Hagenaars, S. P., Davies, G., Wardlaw, J. M., Gale, C. R., Bastin, M. E., and Deary, I. J. (2016). Ageing and brain white matter structure in 3,513 UK Biobank participants. *Nature Communications*, 7(1):13629.
- Craddock, C., Giavasis, S., Clark, D., Shezhad, Z., Pellman, J., Gorgolewski, C. F., Moodie, C. A., and Esteban, O. (2016). PCP: Quality Assessment Protocol. <http://preprocessed-connectomes-project.org/quality-assessment-protocol/>. Accessed: 2016-08-20.
- Croall, I. D., Sanders, D. S., Hadjivassiliou, M., and Hoggard, N. (2020). Cognitive Deficit and White Matter Changes in Persons With Celiac Disease: A Population-Based Study. *Gastroenterology*, 158(8):2112–2122.
- Daducci, A., Canales-Rodríguez, E. J., Zhang, H., Dyrby, T. B., Alexander, D. C., and Thiran, J.-P. (2015). Accelerated Microstructure Imaging via Convex Optimization (AMICO) from diffusion MRI data. *NeuroImage*, 105:32–44.
- Dafflon, J., Pinaya, W. H. L., Turkheimer, F., Cole, J. H., Leech, R., Harris, M. A., Cox, S. R., Whalley, H. C., McIntosh, A. M., and Hellyer, P. J. (2020). An automated machine learning approach to predict brain age from cortical anatomical measures. *Human Brain Mapping*, page hbm.25028.
- Dai, C., Mo, Y., Angelini, E., Guo, Y., and Bai, W. (2019). Transfer learning from partial annotations for whole brain segmentation. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 199–206. Springer.

- Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis I: Segmentation and surface reconstruction. *Neuroimage*, 9:179–94.
- Davenport, S. and Nichols, T. E. (2020). Selective peak inference: Unbiased estimation of raw and standardized effect size at local maxima. *NeuroImage*, 209:116375.
- Daviet, R., Aydogan, G., Jagannathan, K., Spilka, N., Koellinger, P., Kranzler, H. R., Nave, G., and Wetherill, R. R. (2020). Multimodal brain imaging study of 19,825 participants reveals adverse effects of moderate drinking. preprint, Neuroscience.
- de Groot, M., Vernooij, M. W., Klein, S., Ikram, M. A., Vos, F. M., Smith, S., Niessen, W. J., and Andersson, J. L. (2013). Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. *Neuroimage*, 76:400–411.
- de Lange, A.-M. G., Anatürk, M., Kaufmann, T., Cole, J. H., Griffanti, L., Zsoldos, E., Jensen, D., Suri, S., Filippini, N., Singh-Manoux, A., Kivimäki, M., Westlye, L. T., and Ebmeier, K. P. (2020a). Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study. preprint, Neuroscience.
- de Lange, A.-M. G., Barth, C., Kaufmann, T., Anatürk, M., Suri, S., Ebmeier, K. P., and Westlye, L. T. (2020b). History of childbirths relates to region-specific brain aging patterns in middle and older-aged women. preprint, Neuroscience.
- de Lange, A.-M. G., Barth, C., Kaufmann, T., Maximov, I. I., van der Meer, D., Agartz, I., and Westlye, L. T. (2019a). Women’s brain aging: effects of sex-hormone exposure, pregnancies, and genetic risk for Alzheimer’s disease. preprint, Neuroscience.
- de Lange, A.-M. G. and Cole, J. H. (2020). Commentary: Correction procedures in brain-age prediction. *NeuroImage: Clinical*, 26:102229.
- de Lange, A.-M. G., Kaufmann, T., van der Meer, D., Maglanoc, L. A., Alnæs, D., Moberget, T., Douaud, G., Andreassen, O. A., and Westlye, L. T. (2019b). Population-based neuroimaging reveals traces of childbirth in the maternal brain. *Proceedings of the National Academy of Sciences*, 116(44):22341–22346.
- Dekkers, I. A., Jansen, P. R., and Lamb, H. J. (2019). Obesity, Brain Volume, and White Matter Microstructure at MRI: A Cross-sectional UK Biobank Study. *Radiology*, 291(3):763–771.
- Deshmukh, M. and Bhosale, U. (2010). Image fusion and image quality assessment of fused images. *International Journal of Image Processing (IJIP)*, 4(5):484–508.
- 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., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1):1–15.
- Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164. ACM.
- Duff, E. P., Vennart, W., Wise, R. G., Howard, M. A., Harris, R. E., Lee, M., Wartolowska, K., Wani-gasekera, V., Wilson, F. J., Whitlock, M., et al. (2015). Learning to identify CNS drug action and efficacy using multistudy fmri data. *Science translational medicine*, 7(274):274ra16–274ra16.
- Dukart, J., Schroeter, M. L., and Mueller, K. (2011). Age correction in dementia—matching to a healthy brain. *PLoS ONE*, 6(7):e22193.
- Eitel, F., Albrecht, J. P., Paul, F., and Ritter, K. (2019). Harnessing spatial MRI normalization: patch individual filter layers for CNNs. *arXiv:1911.06278 [cs]*.
- Elliott, M. L., Belsky, D. W., Anderson, K., Corcoran, D. L., Ge, T., Knodt, A., Prinz, J. A., Sugden, K., Williams, B., Ireland, D., Poulton, R., Caspi, A., Holmes, A., Moffitt, T., and Hariri, A. R. (2018). A Polygenic Score for Higher Educational Attainment is Associated with Larger Brains. preprint, Neuroscience.
- Elvsashagen, T., Bahrami, S., van der Meer, D., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., Beyer, M. K., Blasi, G., et al. (2019). The genetic architecture of human brainstem structures and their involvement in common brain disorders. *bioRxiv*, page 811711.
- Eshaghi, A., Young, A., Wijertane, P., Prados, F., Arnold, D., Narayanan, S., Guttmann, C. R., Barkhof, F., Alexander, D. C., Thompson, A. J., Chard, D., and Cicarelli, O. (2019). Defining multiple sclerosis phenotypes using MRI. preprint, Neurology.
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J. (2017). Mriqc: Predicting quality in manual mri assessment protocols using no-reference image quality measures. *bioRxiv*, page 111294.
- Esteban, O., Gorgolewski, K. F., A., M. C., and Tripplett, W. (2016). MRIQC: Image quality metrics for quality assessment of MRI. <http://mriqc.readthedocs.io/en/latest/>. Accessed: 2016-08-20.
- Evangelou, E., Gao, H., Chu, C., Ntritsos, G., Blakeley, P., Butts, A. R., Pazoki, R., Suzuki, H., Koskeridis, F., Yiorkas, A. M., et al. (2019). New alcohol-related genes suggest shared genetic mechanisms with neuropsychiatric disorders. *Nature Human Behaviour*, 3(9):950–961.
- Fawns-Ritchie, C. and Deary, I. J. (2019). Reliability and validity of the uk biobank cognitive tests.

medRxiv.

- Filippini, N., MacIntosh, B., Hough, M., Goodwin, G., Frisoni, G., Smith, S., Matthews, P., Beckmann, C., and Mackay, C. (2009). Distinct Patterns of Brain Activity in Young Carriers of the APOE-e4 Allele. *Proc Natl Acad Sci USA (PNAS)*, 106:7209–7214.
- Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T., Mohlberg, H., Amunts, K., and Zilles, K. (2008). Cortical folding patterns and predicting cytoarchitecture. *Cerebral cortex*, 18(8):1973–1980.
- 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.
- Fjell, A. M., Sørensen, o., Amlie, I. K., Bartres-Faz, D., Bros, D. M., Buchmann, N., Demuth, I., Drevon, C. A., Düzel, S., Ebmeier, K. P., Idland, A.-V., Kietzmann, T. C., Kievit, R., Kühn, S., Lindenberger, U., Mowinkel, A. M., Nyberg, L., Price, D., Sexton, C. E., Solé-Padullés, C., Pudas, S., Sederevicius, D., Suri, S., Wagner, G., Watne, L. O., Westerhausen, R., Zsoldos, E., and Walhovd, K. B. (2020). Self-reported sleep relates to hippocampal atrophy across the adult lifespan: results from the Lifebrain consortium. *Sleep*, 43(5):zsz280.
- Flitney, D. E. and Jenkinson, M. (2000). Cluster analysis revisited. In *Tech. rept. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK. TR00DF1*, page 1. Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK.
- Focke, N. K., Helms, G., Kaspar, S., Diederich, C., Tóth, V., Dechent, P., Mohr, A., and Paulus, W. (2011). Multi-site voxel-based morphometry—not quite there yet. *Neuroimage*, 56(3):1164–1170.
- Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.
- Friedman, L. and Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *Journal of Magnetic Resonance Imaging*, 23(6):827–839.
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn Reson Med*, 35(3):346–355.
- Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N. E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology*, 186(9):1026–1034.
- Gale, S. D., Erickson, L. D., Anderson, J. E., Brown, B. L., and Hedges, D. W. (2020). Association between exposure to air pollution and prefrontal cortical volume in adults: A cross-sectional study from the UK biobank. *Environmental Research*, 185:109365.
- García-García, I., Morys, F., and Dagher, A. (2019). Nucleus accumbens volume is related to obesity measures in an age-dependent fashion. *Journal of Neuroendocrinology*.
- Garfield, V., Farmaki, A.-E., Eastwood, S. V., Mathur, R., Rentsch, C. T., Bhaskaran, K., Smeeth, L., and Chaturvedi, N. (2020a). Is glycaemia associated with poorer brain health and risk of dementia? Cross sectional and follow-up analysis of the UK Biobank. preprint, Epidemiology.
- Garfield, V., Farmaki, A.-E., Fatemifar, G., Eastwood, S. V., Mathur, R., Rentsch, C. T., Denaxas, S., Bhaskaran, K. T., Smeeth, L., and Chaturvedi, N. (2020b). The relationship between glycaemia, cognitive function, structural brain outcomes and dementia: A Mendelian randomization study in the UK Biobank. preprint, Genetic and Genomic Medicine.
- Ge, T., Chen, C.-Y., Doyle, A. E., Vettermann, R., Tuominen, L. J., Holt, D. J., Sabuncu, M. R., and Smoller, J. W. (2018). The Shared Genetic Basis of Educational Attainment and Cerebral Cortical Morphology. preprint, Genetics.
- Georgakis, M. K., Malik, R., Anderson, C. D., Parhofer, K. G., Hopewell, J. C., and Dichgans, M. (2020). Genetic determinants of blood lipids and cerebral small vessel disease: role of high-density lipoprotein cholesterol. *Brain*, page awz413.
- Gheorghe, D. A., Li, C., Gallacher, J., and Bauermeister, S. (2019). Associations of adverse lifetime experiences with brain structure in 7003 UK Biobank participants. preprint, Neuroscience.
- Gilmore, A., Buser, N., and Hanson, J. L. (2019). Variations in structural mri quality impact measures of brain anatomy: Relations with age and other sociodemographic variables. *Biorxiv*, page 581876.
- 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., Van Essen, D. C., and Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–24.
- Glastonbury, C. A., Ferlino, M., Nellåker, C., and Lindgren, C. M. (2018). Adjusting for confounding in unsupervised latent representations of images. *arXiv preprint arXiv:1811.06498*.
- Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.-E., Sherif, T., Deelman, E., et al. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in neuroinformatics*, 9:12.
- Glocker, B., Robinson, R., Castro, D. C., Dou, Q., and Konukoglu, E. (2019). Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects. *arXiv:1910.04597 [cs, eess, q-bio]*.

- Glover, G. H., Mueller, B. A., Turner, J. A., Van Erp, T. G., Liu, T. T., Greve, D. N., Voyvodic, J. T., Rasmussen, J., Brown, G. G., Keator, D. B., et al. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *Journal of Magnetic Resonance Imaging*, 36(1):39–54.
- GNC (2014). The German National Cohort: aims, study design and organization. *European Journal of Epidemiology*, 29(5):371.
- Gong, W., Beckmann, C. F., and Smith, S. M. (2020). Phenotype Discovery from Population Brain Imaging. preprint, Neuroscience.
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5:13.
- Gorgolewski, K. J., Alfaro-Almagro, F., Auer, T., Bellec, P., Capotă, M., Chakravarty, M. M., Churchill, N. W., Cohen, A. L., Craddock, R. C., Devenyi, G. A., et al. (2017). Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS computational biology*, 13(3):e1005209.
- 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*, 3:160044.
- Grasby, K. L., Jahanshad, N., Painter, J. N., Colodro-Conde, L., Bralten, J., Hibar, D. P., Lind, P. A., Pizzagalli, F., Ching, C. R., McMahon, M. A. B., et al. (2020). The genetic architecture of the human cerebral cortex. *Science*, 367(6484).
- Gray, J. C., Thompson, M., Bachman, C., Owens, M. M., Murphy, M., and Palmer, R. (2020). Associations of cigarette smoking with gray and white matter in the UK Biobank. *Neuropsychopharmacology*, 45(7):1215–1222.
- Greve, D. and Fischl, B. (2009). "Accurate and robust brain image alignment using boundary-based registration". *NeuroImage*, 48:63–72.
- Greve, D. N., Brown, G. G., Mueller, B. A., Glover, G., and Liu, T. T. (2013). A survey of the sources of noise in fMRI. *Psychometrika*, 78(3):396–416.
- Greve, D. N., Mueller, B. A., Liu, T., Turner, J. A., Voyvodic, J., Yetter, E., Diaz, M., McCarthy, G., Wallace, S., Roach, B. J., et al. (2011). A novel method for quantifying scanner instability in fMRI. *Magnetic resonance in medicine*, 65(4):1053–1061.
- Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M. F., Duff, E. P., Fitzgibbon, S., Westphal, R., Carone, D., Beckmann, C. F., and Smith, S. M. (2017). Hand classification of fMRI ICA noise components. *Neuroimage*, 154:188–205.
- 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., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K. L., and Smith, S. M. (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., Habets, P., Jacobs, H. I., Mengelers, R., Rozendaal, N., Van Os, J., and Marcelis, M. (2012). The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. *PloS one*, 7(6):e38234.
- Gunstad, J., Cohen, R. A., Tate, D. F., Paul, R. H., Poppas, A., Hoth, K., Macgregor, K. L., and Jefferson, A. L. (2005). Blood pressure variability and white matter hyperintensities in older adults with cardiovascular disease. *Blood pressure*, 14(6):353–358.
- Haacke, E., Xu, Y., Cheng, Y., and Reichenbach, J. (2004). Susceptibility-weighted imaging (SWI). *Magnetic Resonance in Medicine*, 52:612–618.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.
- Hamer, M. and Batty, G. D. (2019). Association of body mass index and waist-to-hip ratio with brain structure: UK Biobank study. *Neurology*, 92(6):e594–e600.
- Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., and Weinberger, D. R. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323.
- Harrison, S. J., Bijsterboch, J. D., Segerdahl, A. R., Fitzgibbon, S. P., Farahibozorg, S.-R., Duff, E. P., Smith, S. M., and Woolrich, M. W. (2020). Modelling subject variability in the spatial and temporal characteristics of functional modes. *NeuroImage*, page 117226.
- Harrison, S. J., Woolrich, M. W., Robinson, E. C., Glasser, M. F., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2015). Large-scale probabilistic functional modes from resting state fmri. *NeuroImage*, 109:217–231.
- Hasan, K. M. (2007). A framework for quality control and parameter optimization in diffusion tensor imaging: theoretical analysis and validation. *Magnetic resonance imaging*, 25(8):1196–1202.
- Haworth, S., Mitchell, R., Corbin, L., Wade, K. H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani,

- G., Paternoster, L., Smith, G. D., et al. (2019). Apparent latent structure within the uk biobank sample has implications for epidemiological analysis. *Nature communications*, 10(1):1–9.
- He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., and Yeo, B. T. (2018). Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics. preprint, Neuroscience.
- Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., and Reuter, M. (2020). Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, page 117012.
- Hernández, M., Guerrero, G. D., Cecilia, J. M., García, J. M., Inuggi, A., Jbabdi, S., Behrens, T. E., and Sotiropoulos, S. N. (2013). Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using GPUs. *PloS one*, 8(4):e61892.
- Hernandez-Fernandez, M., Reguly, I., Jbabdi, S., Giles, M., Smith, S., and Sotiropoulos, S. N. (2019). Using gpus to accelerate computational diffusion mri: From microstructure estimation to tractography and connectomes. *NeuroImage*, 188:598–615.
- Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.
- Iglesias, J. E., Insausti, R., Lerma-Usabiaga, G., Bocchetta, M., Van Leemput, K., Greve, D. N., Van der Kouwe, A., Fischl, B., Caballero-Gaudes, C., Paz-Alonso, P. M., et al. (2018). A probabilistic atlas of the human thalamic nuclei combining ex vivo mri and histology. *Neuroimage*, 183:314–326.
- Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634.
- Iglesias, J. E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M., Initiative, A. D. N., et al. (2016). Bayesian longitudinal segmentation of hippocampal substructures in brain mri using subject-specific atlases. *Neuroimage*, 141:542–555.
- Iglesias, J. E., Van Leemput, K., Bhatt, P., Casillas, C., Dutt, S., Schuff, N., Truran-Sacrey, D., Boxer, A., Fischl, B., Initiative, A. D. N., et al. (2015). Bayesian segmentation of brainstem structures in mri. *Neuroimage*, 113:184–195.
- Jager, K., Zoccali, C., Macleod, A., and Dekker, F. (2008). Confounding: what it is and how to deal with it. *Kidney international*, 73(3):256–260.
- Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., Blangero, J., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., et al. (2013). Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA-DTI working group. *Neuroimage*, 81:455–469.
- Jbabdi, S., Sotiropoulos, S. N., Savio, A. M., Graña, M., and 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., Bannister, P., Brady, J., and Smith, S. (2002). Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2):825–841.
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. (2012). FSL. *Neuroimage*, 62(2):782–790.
- Jenkinson, M. and Smith, S. (2001). A Global Optimisation Method for Robust Affine Registration of Brain Images. *Medical Image Analysis*, 5(2):143–156.
- Jezzard, P. and Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic resonance in medicine*, 34(1):65–73.
- John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.
- Johnstone, T., Ores Walsh, K. S., Greischar, L. L., Alexander, A. L., Fox, A. S., Davidson, R. J., and Oakes, T. R. (2006). Motion correction and the use of motion covariates in multiple-subject fMRI analysis. *Hum Brain Mapp*, 27(10):779–788.
- Jonsson, B. A., Bjornsdottir, G., Thorgeirsson, T. E., Ellingsen, L. M., Walters, G. B., Gudbjartsson, D. F., Stefansson, H., Stefansson, K., and Ulfarsson, M. O. (2019). Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10(1):5409.
- Karch, J. D., Filevich, E., Wenger, E., Lisofsky, N., Becker, M., Butler, O., M?rtensson, J., Lindenberg, U., Brandmaier, A. M., and Kühn, S. (2019). Identifying predictors of within-person variance in MRI-based brain volume estimates. *Neuroimage*, 200:575–589.
- Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., et al. (2018). Genetics of brain age suggest an overlap with common brain disorders. *bioRxiv*, page 303164.
- Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., et al. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623.
- Kazemi, K. and Noorzadeh, N. (2014). Quantitative comparison of spm, fsl, and brainsuite for brain mr image segmentation. *Journal of Biomedical Physics and Engineering*, 4(1 Mar).
- Keenan, K. E., Gimbutas, Z., Dienstfrey, A., and Stupic, K. F. (2019). Assessing effects of scanner

- upgrades for clinical studies. *J Magn Reson Imaging*, 50(6):1948–1954.
- Kernbach, J. M., Yeo, B. T. T., Smallwood, J., Margulies, D. S., Thiebaut de Schotten, M., Walter, H., Sabuncu, M. R., Holmes, A. J., Gramfort, A., Varoquaux, G., Thirion, B., and Bzdok, D. (2018). Sub-specialization within default mode nodes characterized in 10,000 UK Biobank participants. *Proceedings of the National Academy of Sciences*, 115(48):12295–12300.
- Kiesow, H., Dunbar, R. I. M., Kable, J. W., Kalenscher, T., Vogeley, K., Schilbach, L., Marquand, A. F., Wiecki, T. V., and Bzdok, D. (2020). 10,000 social brains: Sex differentiation in human brain anatomy. *Science Advances*, 6(12):eaaz1170.
- Kievit, R. A., Fuhrmann, D., Borgeest, G. S., Simpson-Kent, I. L., and Henson, R. N. A. (2018). The neural determinants of age-related changes in fluid intelligence: a pre-registered, longitudinal analysis in UK Biobank. *Wellcome Open Research*, 3:38.
- Kittler, J., Hatef, M., Duin, R. P., and Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239.
- Klapwijk, E. T., van de Kamp, F., van der Meulen, M., Peters, S., and Wierenga, L. M. (2019). Qoala-T: A supervised-learning tool for quality control of FreeSurfer segmented MRI data. *Neuroimage*, 189:116–129.
- Klein, A. and Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171.
- Kolbeinnsson, A., Kossaifi, J., Panagakis, Y., Bulat, A., Anandkumar, A., Tzoulaki, I., and Matthews, P. (2019). Robust Deep Networks with Randomized Tensor Regression Layers. *arXiv:1902.10758 [cs, stat]*.
- Kong, X.-Z., Postema, M., Carrión Castillo, A., Pepe, A., Crivello, F., Joliot, M., Mazoyer, B., Fisher, S. E., and Francks, C. (2019). Handedness and Other Variables Associated with Human Brain Asymmetrical Skew. preprint, Neuroscience.
- Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B. R., Johnson, H., Cash, D., Tabrizi, S. J., Scahill, R. I., Ronneberger, O., and Kl?ppel, S. (2014). Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *Neuroimage*, 98:405–415.
- Krueger, G., Granziera, C., Jack, C. R., Gunter, J. L., Littmann, A., Mortamet, B., Kannengiesser, S., Sorensen, A. G., Ward, C. P., Reyes, D. A., Britson, P. J., Fischer, H., and Bernstein, M. A. (2012). Effects of MRI scan acceleration on brain volume measurement consistency. *J Magn Reson Imaging*, 36(5):1234–1240.
- Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *NeuroImage*, 169:431–442.
- Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.
- Lange, F. J., Ashburner, J., Smith, S. M., and Andersson, J. L. (2020). A symmetric prior for the regularisation of elastic deformations: Improved anatomical plausibility in nonlinear image registration. *NeuroImage*, page 116962.
- Le Guen, Y., Karkar, S., Grigis, A., Philippe, C., Mangin, J.-F., and Frouin, V. (2019a). Heritability of surface area and cortical thickness: a comparison between the Human Connectome Project and the UK Biobank dataset. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1887–1890, Venice, Italy. IEEE.
- Le Guen, Y., Leroy, F., Philippe, C., Mangin, J.-F., Dehaene-Lambertz, G., and Frouin, V. (2020). Enhancer Locus in ch14q23.1 Modulates Brain Asymmetric Temporal Regions Involved in Language Processing. *bioRxiv preprint* <https://www.biorxiv.org/content/10.1101/539189v2>.
- Le Guen, Y., Philippe, C., Riviere, D., Lemaitre, H., Grigis, A., Fischer, C., Dehaene-Lambertz, G., Mangin, J.-F., and Frouin, V. (2019b). eQTL of KCNK2 regionally influences the brain sulcal widening: evidence from 15,597 UK Biobank participants with neuroimaging data. *Brain Structure and Function*, 224(2):847–857.
- Lee, J. J., McGue, M., Iacono, W. G., Michael, A. M., and Chabris, C. F. (2019). The causal influence of brain size on human intelligence: Evidence from within-family phenotypic associations and GWAS modeling. *Intelligence*, 75:48–58.
- Leming, M. and Suckling, J. (2020). Stochastic encoding of graphs in deep learning allows for complex analysis of gender classification in resting-state and task functional brain networks from the UK Biobank. *arXiv:2002.10936 [cs, q-bio, stat]*.
- Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264:47–56.
- Liao, D., Cooper, L., Cai, J., Toole, J. F., Bryan, N. R., Hutchinson, R. G., and Tyroler, H. A. (1996). Presence and severity of cerebral white matter lesions and hypertension, its treatment, and its control. *Stroke*, 27(12):2262–2270.
- Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J. D., Boulton, C., Collins, R., Conroy, M. C., et al. (2020). The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11(1):1–12.
- Liu, J., Rutten-Jacobs, L., Liu, M., Markus, H. S., and Traylor, M. (2018). Causal Impact of Type 2 Diabetes Mellitus on Cerebral Small Vessel Disease: A Mendelian Randomization Analysis. *Stroke*,

- 49(6):1325–1331.
- Liu, Z., Wang, Y., Gerig, G., Gouttard, S., Tao, R., Fletcher, T., and Styner, M. (2010). Quality control of diffusion weighted images. In *SPIE medical imaging*, pages 76280J–76280J. International Society for Optics and Photonics.
- Liu, Z., Xiao, X., Zhang, K., Zhao, Q., Cao, X., Li, C., Wang, M., Lin, W., Qiu, J., Sahakian, B. J., Feng, J., Robbins, T. W., and Zhang, J. (2019). Dynamic brain network changes in resting-state reflect neuroplasticity: molecular and cognitive evidence. preprint, Neuroscience.
- Lund, M. J., Alnæs, D., Schwab, S., Meer, D., Andreassen, O. A., Westlye, L. T., and Kaufmann, T. (2020). Differences in directed functional brain connectivity related to age, sex and mental health. *Human Brain Mapping*, page hbm.25116.
- Luo, N., Sui, J., Abrol, A., Chen, J., Turner, J. A., Damaraju, E., Fu, Z., Fan, L., Lin, D., Zhuo, C., Xu, Y., Glahn, D. C., Rodrigue, A. L., Banich, M. T., Pearlson, G. D., and Calhoun, V. D. (2019). Structural Brain Architectures Match Intrinsic Functional Networks and Vary across Domains: A Study from 15 000+ Individuals. *Cerebral Cortex*, page 11.
- Lyall, D. M., Cox, S. R., Lyall, L. M., Celis-Morales, C., Cullen, B., Mackay, D. F., Ward, J., Strawbridge, R. J., McIntosh, A. M., Sattar, N., Smith, D. J., Cavanagh, J., Deary, I. J., and Pell, J. P. (2019). Association between APOE e4 and white matter hyperintensity volume, but not total brain volume or white matter integrity. *Brain Imaging and Behavior*.
- MacCarthy, P. (2017). fslpy. <https://git.fmrib.ox.ac.uk/fsl/fslpy>.
- MacCarthy, P. (2019). Funpack (version 1.4.1). Zenodo: <http://doi.org/10.5281/zenodo.3371332>. Accessed: 2019-12-10.
- Maglanoc, L. A., Kaufmann, T., van der Meer, D., Marquand, A. F., Wolfers, T., Jonassen, R., Hilland, E., Andreassen, O. A., Landrø, N. I., and Westlye, L. T. (2019). Predicting cognitive and mental health traits and their polygenic architecture using large-scale brain connectomics. preprint, Neuroscience.
- Majrashi, N. A., Ahearn, T. S., and Waiter, G. D. (2020a). Brainstem volume mediates seasonal variation in depressive symptoms: A cross sectional study in the UK Biobank cohort. *Scientific Reports*, 10(1):3592.
- Majrashi, N. A., Ahearn, T. S., Williams, J. H. G., and Waiter, G. D. (2020b). Sex differences in the association of photoperiod with hippocampal subfield volumes in older adults: A cross-sectional study in the UK Biobank cohort. *Brain and Behavior*, 10(6).
- Mandal, A. S., Romero-Garcia, R., Hart, M., and Suckling, J. (2020). Genetic, Cellular, and Connectomic Characterization of the Adult Human Brain Regions Commonly Plagued by Glioma. preprint, Neurology.
- Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., Barch, D. M., Archie, K. A., Burgess, G. C., Ramaratnam, M., et al. (2013). Human Connectome Project informatics: quality control, database services, and data visualization. *Neuroimage*, 80:202–219.
- Markiewicz, C. J., Appelhoff, S., Calhoun, V., Dickie, E. W., Duff, E., DuPre, E., Esteban, O., Feingold, F., Ghosh, S., Halchenko, Y. O., Harms, M. P., Herholz, P., Mennes, M., Nørgaard, M., Oostenveld, R., Pernet, C., Pestilli, F., Poldrack, R. A., Rokem, A., Smith, R. E., Yarkoni, T., and Gorgolewski, K. J. (2020). Bids derivatives: Standardization of processing results in brain imaging. *Organization for Human Brain Mapping*, page 1827.
- Maximov, I. I., Alnæs, D., and Westlye, L. T. (2019a). Towards an optimised processing pipeline for diffusion magnetic resonance imaging data: Effects of artefact corrections on diffusion metrics and their age associations in uk biobank. *Human brain mapping*, 40(14):4146–4162.
- Maximov, I. I., Alnæs, D., and Westlye, L. T. (2019b). Towards an optimised processing pipeline for diffusion magnetic resonance imaging data: Effects of artefact corrections on diffusion metrics and their age associations in UK Biobank. *Human Brain Mapping*, 40(14):4146–4162.
- Maximov, I. I., van der Meer, D., de Lange, A.-M., Kaufmann, T., Shadrin, A., Frei, O., Wolfers, T., and Westlye, L. T. (2020). Fast quality control method for derived diffusion metrics (YTTRIUM) in big data analysis: UK Biobank 18608 example. preprint, Neuroscience.
- Maximov, I. I. and Westlye, L. T. (2020). Feasibility of axon density metrics for brain asymmetry evaluation in the UK Biobank subsample. preprint, Neuroscience.
- 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., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M., and Smith, S. M. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.*, 19(11):1523–1536.
- Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., and Ugurbil, K. (2010). Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic Resonance in Medicine*, 63(5):1144–1153.
- Mollink, J., Smith, S. M., Elliott, L. T., Kleinnijenhuis, M., Hiemstra, M., Alfaro-Almagro, F., Marchini, J., van Cappellen van Walsum, A.-M., Jbabdi, S., and Miller, K. L. (2019). The spatial correspondence and genetic influence of interhemispheric connectivity with white matter microstructure. *Nature Neuroscience*, 22(5):809–819.

- Moorthy, A. K. and Bovik, A. C. (2010). A two-step framework for constructing blind image quality indices. *Signal Processing Letters, IEEE*, 17(5):513–516.
- Moreau, C., Huguet, G., Urchs, S., Douard, E. A., Sharmarke, H., Orban, P., Labbe, A., Modenato, C., Martin-Brevet, S., Kuldeep, K., et al. (2020). The general impact of haploinsufficiency on brain connectivity underlies the pleiotropic effect of neuropsychiatric cnvs. Technical report, Genetic and Genomic Medicine.
- Mori, S., Wakana, S., Van Zijl, P. C., and Nagae-Poetscher, L. (2005). *MRI atlas of human white matter*, volume 16. Am Soc Neuroradiology.
- Mortamet, B., Bernstein, M. A., Jack, C. R., Gunter, J. L., Ward, C., Britson, P. J., Meuli, R., Thiran, J.-P., and Krueger, G. (2009). Automatic Quality Assessment in structural brain Magnetic Resonance Imaging. *Magnetic Resonance in Medicine*, 62(2):365–372.
- Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2005). Ways toward an early diagnosis in alzheimer’s disease: the alzheimer’s disease neuroimaging initiative (adni). *Alzheimer’s & Dementia*, 1(1):55–66.
- Murphy, K., Birn, R. M., and Bandettini, P. A. (2013). Resting-state fMRI confounds and cleanup. *Neuroimage*, 80:349–359.
- Murphy, K. and Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage*, 154:169–173.
- Nave, G., Jung, W. H., Karlsson Linnér, R., Kable, J. W., and Koellinger, P. D. (2019). Are bigger brains smarter? evidence from a large-scale preregistered study. *Psychological science*, 30(1):43–54.
- Neilson, E., Shen, X., Cox, S. R., Clarke, T.-K., Wigmore, E. M., Gibson, J., Howard, D. M., Adams, M. J., Harris, M. A., Davies, G., Deary, I. J., Whalley, H. C., McIntosh, A. M., and Lawrie, S. M. (2019). Impact of Polygenic Risk for Schizophrenia on Cortical Structure in UK Biobank. *Biological Psychiatry*, 86(7):536–544.
- Nichols, T. (2013). *Notes on creating a standardized version of DVARS*.
- Ning, K., Zhao, L., Franklin, M., Matloff, W., Batta, I., Arzouni, N., Sun, F., and Toga, A. W. (2020a). Parity is associated with cognitive function and brain age in both females and males. *Scientific Reports*, 10(1):6100.
- Ning, K., Zhao, L., Matloff, W., Sun, F., and Toga, A. W. (2020b). Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants. *Scientific Reports*, 10(1):10.
- Nobis, L., Manohar, S. G., Smith, S. M., Alfaro-Almagro, F., Jenkinson, M., Mackay, C. E., and Husain, M. (2019). Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical*, 23:101904.
- Noble, S., Scheinost, D., Finn, E. S., Shen, X., Papademetris, X., McEwen, S. C., Bearden, C. E., Addington, J., Goodyear, B., Cadenhead, K. S., Mirzakhani, H., Cornblatt, B. A., Olvet, D. M., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., Belger, A., Seidman, L. J., Thermenos, H., Tsuang, M. T., van Erp, T. G. M., Walker, E. F., Hamann, S., Woods, S. W., Cannon, T. D., and Constable, R. T. (2017). Multisite reliability of MR-based functional connectivity. *Neuroimage*, 146:959–970.
- Norbury, R. (2019). Chronotype, depression and hippocampal volume: cross-sectional associations from the UK Biobank. *Chronobiology International*, 36(5):709–716.
- Norbury, R. (2020). Diurnal Preference and Grey Matter Volume in a Large Population of Older Adults: Data from the UK Biobank. *Journal of Circadian Rhythms*, 18(1):3.
- O’Connell, K. S., Sønderby, I. E., Frei, O., van der Meer, D., Athanasiu, L., Smeland, O. B., Alnæs, D., Kaufmann, T., Westlye, L. T., Steen, V. M., Andreassen, O., Hughes, T., and Djurovic, S. (2020). Association between complement component 4A expression, cognitive performance and brain imaging measures in UK Biobank. preprint, Genetic and Genomic Medicine.
- Park, B.-y., Byeon, K., Lee, M. J., Kim, S.-H., and Park, H. (2020). The orbitofrontal cortex functionally links obesity and white matter hyperintensities. *Scientific Reports*, 10(1):2930.
- Parker Jones, O., Alfaro-Almagro, F., and Jbabdi, S. (2018). An empirical, 21st century evaluation of phrenology. *Cortex*, 106:26–35.
- Patenaude, B., Smith, S., Kennedy, D., and Jenkinson, M. (2011). A Bayesian Model of Shape and Appearance for Subcortical Brain Segmentation. *NeuroImage*, 56(3):907–922.
- Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., and Smith, S. M. (2019). Accurate brain age prediction with lightweight deep neural networks. preprint, Neuroscience.
- Persyn, E., Hanscombe, K. B., Howson, J. M. M., Lewis, C. M., Traylor, M., and Markus, H. S. (2020). Genome-wide association study of MRI markers of cerebral small vessel disease in 42,310 participants. *Nature Communications*, 11(1):2175.
- Pervaiz, U., Vidaurre, D., Woolrich, M. W., and Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, 211:116604.
- Pinaya, W. H. L., Scarpazza, C., Garcia-Dias, R., Vieira, S., Baecker, L., da Costa, P. F., Redolfi, A., Frisoni, G. B., Pievani, M., Calhoun, V. D., Sato, J. R., Mechelli, A., the Alzheimer’s Disease

- Neuroimaging Initiative, and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (2020). Normative modelling using deep autoencoders: a multi-cohort study on mild cognitive impairment and Alzheimer's disease. preprint, Neuroscience.
- Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Wolf, D. H., Gur, R., Gur, R., Morris, J., Albert, M. S., Grabe, H. J., Resnick, S. M., Bryan, R. N., Wolk, D. A., Shinohara, R. T., Shou, H., and Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450.
- Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, 59(3):2142–2154.
- Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *Neuroimage*, 84:320–341.
- Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P. M., and Glocker, B. (2018). Neuronet: fast and robust reproduction of multiple brain image segmentation pipelines. *arXiv preprint arXiv:1806.04224*.
- Rao, A., Monteiro, J. M., and Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage*, 150:23–49.
- Reus, L. M., Shen, X., Gibson, J., Wigmore, E., Ligthart, L., Adams, M. J., Davies, G., Cox, S. R., Hagenaars, S. P., Bastin, M. E., Deary, I. J., Whalley, H. C., and McIntosh, A. M. (2017). Association of polygenic risk for major psychiatric illness with subcortical volumes and white matter integrity in UK Biobank. *Scientific Reports*, 7(1):42140.
- Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., Harris, M. A., Alderson, H. L., Hunter, S., Neilson, E., et al. (2018). Sex differences in the adult human brain: evidence from 5216 uk biobank participants. *Cerebral Cortex*, 28(8):2959–2975.
- Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J. L., Burgess, G. C., Harms, M. P., Smith, S., Van Essen, D. C., and Jenkinson, M. (2014). Msm: A new flexible framework for multimodal surface matching. *Neuroimage*, 100:414–426.
- Rodrigue, A. L., Alexander-Bloch, A. F., Knowles, E. E. M., Mathias, S. R., Mollon, J., Koenis, M. M. G., Perrone-Bizzozero, N. I., Almasy, L., Turner, J. A., Calhoun, V. D., and Glahn, D. C. (2020). Genetic Contributions to Multivariate Data-Driven Brain Networks Constructed via Source-Based Morphometry. *Cerebral Cortex*, page bhaa082.
- Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., et al. (2018). Quantitative assessment of structural image quality. *Neuroimage*, 169:407–418.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rudner, M., Seeto, M., Keidser, G., Johnson, B., and Rönnerberg, J. (2019). Poorer Speech Reception Threshold in Noise Is Associated With Lower Brain Volume in Auditory and Cognitive Processing Regions. *Journal of Speech, Language, and Hearing Research*, 62(4S):1117–1130.
- Rutten-Jacobs, L. C., Tozer, D. J., Duering, M., Malik, R., Dichgans, M., Markus, H. S., and Traylor, M. (2018). Genetic Study of White Matter Integrity in UK Biobank (N=8448) and the Overlap With Stroke, Depression, and Dementia. *Stroke*, 49(6):1340–1347.
- Salat, D. H., Lee, S. Y., Van der Kouwe, A., Greve, D. N., Fischl, B., and Rosas, H. D. (2009). Age-associated alterations in cortical gray and white matter signal intensity and gray to white matter contrast. *Neuroimage*, 48(1):21–28.
- Salimi-Khorshidi, G., Douaud, G., Beckmann, C., Glasser, M., Griffanti, L., and Smith, S. (2014). Automatic Denoising of Functional MRI Data: Combining Independent Component Analysis and Hierarchical Fusion of Classifiers. *NeuroImage*, 90:449–468.
- Salminen, L. E., Wilcox, R. R., Zhu, A. H., Riedel, B. C., Ching, C. R. K., Rashid, F., Thomopoulos, S. I., Saremi, A., Harrison, M. B., Ragothaman, A., Knight, V., Boyle, C. P., Medland, S. E., Thompson, P. M., and Jahanshad, N. (2018). Altered cortical brain structure and increased risk for disease seen decades after perinatal exposure to maternal smoking: A study of 9,000 adults in the UK Biobank. preprint, Neuroscience.
- Santana, A. N., Cifre, I., de Santana, C. N., and Montoya, P. (2019). Using Deep Learning and Resting-State fMRI to Classify Chronic Pain Conditions. *Frontiers in Neuroscience*, 13:1313.
- Satizabal, C. L., Adams, H. H., Hibar, D. P., White, C. C., Knol, M. J., Stein, J. L., Scholz, M., Sargurupremraj, M., Jahanshad, N., Roshchupkin, G. V., et al. (2019). Genetic architecture of subcortical brain structures in 38,851 individuals. *Nature genetics*, 51(11):1624–1636.
- Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughhead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, 64:240–256.
- Savalia, N. K., Agres, P. F., Chan, M. Y., Feczko, E. J., Kennedy, K. M., and Wig, G. S. (2017). Motion-

- related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. *Hum Brain Mapp*, 38(1):472–492.
- Saygin, Z. M., Kliemann, D., Iglesias, J. E., van der Kouwe, A. J., Boyd, E., Reuter, M., Stevens, A., Van Leemput, K., McKee, A., Frosch, M. P., et al. (2017). High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. *Neuroimage*, 155:370–382.
- Schram, M. T., Sep, S. J., van der Kallen, C. J., Dagnelie, P. C., Koster, A., Schaper, N., Henry, R. M., and Stehouwer, C. D. (2014). The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451.
- Schulz, M.-A., Chapman-Rounds, M., Verma, M., Bzdok, D., and Georgatzis, K. (2020). Clusters in Explanation Space: Inferring disease subtypes from model explanations. *arXiv:1912.08755 [cs, q-bio, stat]*.
- Schulz, M.-A., Yeo, B. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., and Bzdok, D. (2019). Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets. preprint, Bioinformatics.
- Ségonne, F., Dale, A., Busa, E., Glessner, M., Salat, D., Hahn, H., and Fischl, B. (2004). A hybrid approach to the skull stripping problem in mri. *Neuroimage*, 22(3):1060–1075.
- Sha, Z., Schijven, D., Carrion-Castillo, A., Joliot, M., Mazoyer, B., Fisher, S. E., Crivello, F., and Francks, C. (2020). The genetic architecture of structural left-right asymmetry of the human brain. preprint, Neuroscience.
- Shattuck, D. W. and Leahy, R. M. (2002). Brainsuite: an automated cortical surface identification tool. *Medical image analysis*, 6(2):129–142.
- Shen, J., Tozer, D. J., Markus, H. S., and Tay, J. (2020). Network Efficiency Mediates the Relationship Between Vascular Burden and Cognitive Impairment: A Diffusion Tensor Imaging Study in UK Biobank. *Stroke*, 51(6):1682–1689.
- Shen, X., Adams, M. J., Ritakari, T. E., Cox, S. R., McIntosh, A. M., and Whalley, H. C. (2019). White Matter Microstructure and Its Relation to Longitudinal Measures of Depressive Symptoms in Mid- and Late Life. *Biological Psychiatry*, 86(10):759–768.
- Shen, X., Cox, S. R., Adams, M. J., Howard, D. M., Lawrie, S. M., Ritchie, S. J., Bastin, M. E., Deary, I. J., McIntosh, A. M., and Whalley, H. C. (2018). Resting-State Connectivity and Its Association With Cognitive Performance, Educational Attainment, and Household Income in the UK Biobank. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(10):878–886.
- Shen, X., Reus, L. M., Cox, S. R., Adams, M. J., Liewald, D. C., Bastin, M. E., Smith, D. J., Deary, I. J., Whalley, H. C., and McIntosh, A. M. (2017). Subcortical volume and white matter integrity abnormalities in major depressive disorder: findings from UK Biobank imaging data. *Scientific Reports*, 7(1):5547.
- Smith, S. (2002). Fast Robust Automated Brain Extraction. *Human Brain Mapping*, 17(3):143–155.
- Smith, S. (2012). The Future of fMRI Connectivity. *NeuroImage*, 62:1257–1266.
- Smith, S., Beckmann, C. F., Andersson, J. L., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., et al. (2013). Resting-state fMRI in the human connectome project. *Neuroimage*, 80:144–168.
- Smith, S., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain’s functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045.
- Smith, S., Hyvärinen, A., Varoquaux, G., Miller, K., and Beckmann, C. (2014). Group-PCA for very large fMRI datasets. *NeuroImage*, 101:738–749.
- Smith, S., 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., Elliott, L. T., Alfaro-Almagro, F., McCarthy, P., Nichols, T. E., Douaud, G., and Miller, K. L. (2020). Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *eLife*, 9:e52677.
- Smith, S. M. and Nichols, T. E. (2018). Statistical challenges in “big data” human neuroimaging. *Neuron*, 97(2):263–268.
- Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., and Miller, K. L. (2019). Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539.
- Smith, S. M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P. M., Federico, A., and De Stefano, N. (2002). Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage*, 17(1):479–489.
- Snoek, L., Mileti?, S., and Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *Neuroimage*, 184:741–760.
- Soheili-Nezhad, S., Jahanshad, N., Guelfi, S., Khosrowabadi, R., Saykin, A. J., Thompson, P. M., Beckmann, C. F., Sprooten, E., Zarei, M., and the Alzheimer’s Disease Neuroimaging (2020). Imaging genomics discovery of a new risk variant for Alzheimer’s disease in the postsynaptic <span style=

- variant:small-caps;">SHARPIN gene. *Human Brain Mapping*, page hbm.25083.
- Song, X.-W., Dong, Z.-Y., Long, X.-Y., Li, S.-F., Zuo, X.-N., Zhu, C.-Z., He, Y., Yan, C.-G., and Zang, Y.-F. (2011). REST: a toolkit for resting-state functional magnetic resonance imaging data processing. *PloS one*, 6(9):e25031.
- Stankevičiūtė, K., Azevedo, T., Campbell, A., Bethlehem, R., and Liò, P. (2020). Population Graph GNNs for Brain Age Prediction. preprint, Neuroscience.
- Stein, J. L., Medland, S. E., Vasquez, A. A., Hibar, D. P., Senstad, R. E., Winkler, A. M., Toro, R., Appel, K., Bartecek, R., Bergmann, Ø., et al. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nature genetics*, 44(5):552–561.
- Stoffers, D., Diaz, B. A., Chen, G., den Braber, A., van 't Ent, D., Boomsma, D. I., Mansvelder, H. D., de Geus, E., Van Someren, E. J., and Linkenkaer-Hansen, K. (2015). Resting-State fMRI Functional Connectivity Is Associated with Sleepiness, Imagery, and Discontinuity of Mind. *PLoS ONE*, 10(11):e0142014.
- Stolicyn, A., Harris, M. A., Shen, X., Barbu, M. C., Adams, M. J., Hawkins, E. L., Nooij, L., Yeung, H. W., Murray, A. D., Lawrie, S. M., Steele, J. D., McIntosh, A. M., and Whalley, H. C. (2020). Automated classification of depression from structural brain measures across two independent community-based cohorts. *Human Brain Mapping*, page hbm.25095.
- 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):1–10.
- Sundaresan, V., Griffanti, L., Kindalova, P., Alfaro-Almagro, F., Zamboni, G., Rothwell, P. M., Nichols, T. E., and Jenkinson, M. (2019a). Modelling the distribution of white matter hyperintensities due to ageing on mri images using bayesian inference. *Neuroimage*, 185:434–445.
- Sundaresan, V., Zamboni, G., Le Heron, C., Rothwell, P. M., Husain, M., Battaglini, M., De Stefano, N., Jenkinson, M., and Griffanti, L. (2019b). Automated lesion segmentation with bianca: impact of population-level features, classification algorithm and locally adaptive thresholding. *NeuroImage*, 202:116056.
- Suzuki, H., Gao, H., Bai, W., Evangelou, E., Glocker, B., O'Regan, D. P., Elliott, P., and Matthews, P. M. (2017). Abnormal brain white matter microstructure is associated with both pre-hypertension and hypertension. *PLoS ONE*, 12(11):e0187600.
- Suzuki, H., Venkataraman, A. V., Bai, W., Guitton, F., Guo, Y., Dehghan, A., Matthews, P. M., and for the Alzheimer's Disease Neuroimaging Initiative (2019). Associations of Regional Brain Structural Differences With Aging, Modifiable Risk Factors for Dementia, and Cognitive Performance. *JAMA Network Open*, 2(12):e1917257.
- Swanson, J. M. (2012). The uk biobank and selection bias. *The Lancet*, 380(9837):110.
- Taebi, A., Kiesow, H., Vogeley, K., Schilbach, L., Bernhardt, B. C., and Bzdok, D. (2020). Population variability in social brain morphology for social support, household size, and friendship satisfaction. *Social Cognitive and Affective Neuroscience*, page nsaa075.
- Taquet, M., Smith, S. M., Prohl, A. K., Peters, J. M., Warfield, S. K., Scherrer, B., and Harrison, P. J. (2020). A structural brain network of genetic vulnerability to psychiatric illness. *Molecular Psychiatry*.
- Traylor, M., Tozer, D. J., Croall, I. D., Lisiecka Ford, D. M., Olorunda, A. O., Boncoraglio, G., Dichgans, M., Lemmens, R., Rosand, J., Rost, N. S., Rothwell, P. M., Sudlow, C. L., Thijs, V., Rutten-Jacobs, L., and Markus, H. S. (2019). Genetic variation in *PLEKHG1* is associated with white matter hyperintensities (n = 11,226). *Neurology*, 92(8):e749–e757.
- Tsang, R. S. M., Gallacher, J. E. J., and Bauermeister, S. (2019). Sex differences in interrelationships between arterial stiffness, carotid intima-media thickness, white matter hyperintensities, depression and cognition: A UK Biobank study. preprint, Neuroscience.
- Vakli, P., Deák-Meszlényi, R. J., Auer, T., and Vidnyánszky, Z. (2020). Predicting Body Mass Index From Structural MRI Brain Images Using a Deep Convolutional Neural Network. *Frontiers in Neuroinformatics*, 14:10.
- van Der Lee, S. J., Knol, M. J., Chauhan, G., Satizabal, C. L., Smith, A. V., Hofer, E., Bis, J. C., Hibar, D. P., Hilal, S., van den Akker, E. B., et al. (2019). A genome-wide association study identifies genetic loci associated with specific lobar brain volumes. *Communications biology*, 2(1):1–9.
- van der Meer, D., Frei, O., Kaufmann, T., Chen, C.-H., Thompson, W., O'Connell, K., Monereo Sánchez, J., Linden, D., Westlye, L., Dale, A. M., and Andreassen, O. (2020). Quantifying the Polygenic Architecture of the Human Cerebral Cortex. *Biological Psychiatry*, 87(9):S131–S132.
- Van der Meer, D., Rokicki, J., Kaufmann, T., Córdova-Palomera, A., Moberget, T., Alnæs, D., Bettella, F., Frei, O., Doan, N. T., Sønderby, I. E., et al. (2018). Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. Technical report, Neuroscience.
- Van Essen, D. C., Smith, S., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79.
- Veldsman, M., Kindalova, P., Husain, M., Kosmidis, I., and Nichols, T. E. (2020). Spatial distribution and cognitive impact of cerebrovascular risk-related white matter hyperintensities. preprint, Neuroscience.
- Veraart, J., Novikov, D. S., Christiaens, D., Ades-Aron, B., Sijbers, J., and Fieremans, E. (2016). Denoising of diffusion mri using random matrix theory. *Neuroimage*, 142:394–406.

- Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A. J., Alfaro-Almagro, F., Smith, S. M., and Woolrich, M. W. (2018). Discovering dynamic brain networks from big data in rest and task. *NeuroImage*, 180:646–656.
- Vidaurre, D., Quinn, A. J., Baker, A. P., Dupret, D., Tejero-Cantero, A., and Woolrich, M. W. (2016). Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage*, 126:81–95.
- Vinke, E. J., Huizinga, W., Bergtholdt, M., Adams, H. H., Steketee, R. M., Papma, J. M., de Jong, F. J., Niessen, W. J., Ikram, M. A., Wenzel, F., and Vernooij, M. W. (2019). Normative brain volumetry derived from different reference populations: impact on single-subject diagnostic assessment in dementia. *Neurobiology of Aging*, 84:9–16.
- Wachinger, C., Rieckmann, A., and Pölsterl, S. (2020). Detect and Correct Bias in Multi-Site Neuroimaging Datasets. *arXiv:2002.05049 [cs, eess]*.
- 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.
- Ward, J., Lyall, L. M., Bethlehem, R. A. I., Ferguson, A., Strawbridge, R. J., Lyall, D. M., Cullen, B., Graham, N., Johnston, K. J. A., Bailey, M. E. S., Murray, G. K., and Smith, D. J. (2019). Novel genome-wide associations for anhedonia, genetic correlation with psychiatric disorders, and polygenic association with brain structure. *Translational Psychiatry*, 9(1):327.
- Wardlaw, J. M., Bastin, M. E., Valdés Hernández, M. C., Maniega, S. M., Royle, N. A., Morris, Z., Clayden, J. D., Sandeman, E. M., Eadie, E., Murray, C., et al. (2011). Brain aging, cognition in youth and old age and vascular disease in the lothian birth cohort 1936: rationale, design and methodology of the imaging protocol. *International Journal of Stroke*, 6(6):547–559.
- Warland, A., Kendall, K. M., Rees, E., Kirov, G., and Caseras, X. (2020). Schizophrenia-associated genomic copy number variants and subcortical brain volumes in the UK Biobank. *Molecular Psychiatry*, 25(4):854–862.
- Warrington, S., Bryant, K., Khrapitchev, A., Sallet, J., Charquero-Ballester, M., Douaud, G., Jbabdi, S., Mars, R., and Sotiropoulos, S. (2019). XTRACT - Standardised protocols for automated tractography in the human and macaque brain. preprint, Neuroscience.
- Warrington, S., Bryant, K. L., Khrapitchev, A. A., Sallet, J., Charquero-Ballester, M., Douaud, G., Jbabdi, S., Mars, R. B., and Sotiropoulos, S. N. (2020). Xtract-standardised protocols for automated tractography in the human and macaque brain. *NeuroImage*, page 116923.
- Wei, X., Warfield, S. K., Zou, K. H., Wu, Y., Li, X., Guimond, A., Mugler, J. P., Benson, R. R., Wolfson, L., Weiner, H. L., et al. (2002). Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. *Journal of Magnetic Resonance Imaging*, 15(2):203–209.
- Wei, Y., de Lange, S. C., Scholtens, L. H., Watanabe, K., Ardesch, D. J., Jansen, P. R., Savage, J. E., Li, L., Preuss, T. M., Rilling, J. K., Posthuma, D., and van den Heuvel, M. P. (2019). Genetic mapping and evolutionary analysis of human-expanded cognitive networks. *Nature Communications*, 10(1):4839.
- Westfall, J. and Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLoS ONE*, 11(3):e0152719.
- Wiberg, A., Ng, M., Al Omran, Y., Alfaro-Almagro, F., McCarthy, P., Marchini, J., Bennett, D. L., Smith, S., Douaud, G., and Furniss, D. (2019). Handedness, language areas and neuropsychiatric diseases: insights from brain imaging and genetics. *Brain*, 142(10):2938–2947.
- Wigmore, E. M., Clarke, T.-K., Howard, D. M., Adams, M. J., Hall, L. S., Zeng, Y., Gibson, J., Davies, G., Fernandez-Pujals, A. M., Thomson, P. A., Hayward, C., Smith, B. H., Hocking, L. J., Padmanabhan, S., Deary, I. J., Porteous, D. J., Nicodemus, K. K., and McIntosh, A. M. (2017). Do regional brain volumes and major depressive disorder share genetic architecture? A study of Generation Scotland (n=19 762), UK Biobank (n=24 048) and the English Longitudinal Study of Ageing (n=5766). *Translational Psychiatry*, 7(8):e1205–e1205.
- Woodard, J. P. and Carley-Spencer, M. P. (2006). No-reference image quality metrics for structural MRI. *Neuroinformatics*, 4(3):243–262.
- Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*, 14(6):1370–1386.
- Wu, C. (2019). Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank. preprint, Genetics.
- Zarnani, K., Nichols, T. E., Alfaro-Almagro, F., Fagerlund, B., Lauritzen, M., Rostrup, E., and Smith, S. M. (2019). Discovering markers of healthy aging: a prospective study in a danish male birth cohort. *Aging (Albany NY)*, 11(16):5943.
- Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*, 61(4):1000–1016.
- Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896.
- Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of Brain MR Images through a Hidden Markov Random Field Model and the Expectation Maximization Algorithm. *IEEE Trans. on Medical Imaging*,

- 20(1):45–57.
- Zhao, B., Ibrahim, J. G., Li, Y., Li, T., Wang, Y., Shan, Y., Zhu, Z., Zhou, F., Zhang, J., Huang, C., et al. (2019a). Heritability of regional brain volumes in large-scale neuroimaging and genetic studies. *Cerebral Cortex*, 29(7):2904–2914.
- Zhao, B., Li, T., Yang, Y., Wang, X., Luo, T., Shan, Y., Zhu, Z., Xiong, D., Hauberg, M. E., Bendl, J., Fullard, J. F., Roussos, P., Li, Y., Stein, J. L., and Zhu, H. (2020). Common genetic variation influencing human white matter microstructure. preprint, Genetics.
- Zhao, B., Zhang, J., Ibrahim, J. G., Luo, T., Santelli, R. C., Li, Y., Li, T., Shan, Y., Zhu, Z., Zhou, F., Liao, H., Nichols, T. E., and Zhu, H. (2019b). Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n = 17,706). *Molecular Psychiatry*.
- Zhou, Y., Zhao, L., Zhou, N., Zhao, Y., Marino, S., Wang, T., Sun, H., Toga, A. W., and Dinov, I. D. (2019). Predictive Big Data Analytics using the UK Biobank Data. *Scientific Reports*, 9(1):6012.
- Zonneveld, H. I., Pruim, R. H., Bos, D., Vrooman, H. A., Muetzel, R. L., Hofman, A., Rombouts, S. A., van der Lugt, A., Niessen, W. J., Ikram, M. A., and Vernooij, M. W. (2019). Patterns of functional connectivity in an aging population: The Rotterdam Study. *NeuroImage*, 189:432–444.