Towards multicentre application of diffusion MRI in cerebral small vessel disease

Bruno M. de Brito Robalo

Towards multicentre application of diffusion MRI in cerebral small vessel disease

Cover design Bruno Robalo
Layout Bruno Robalo
Printed by Ridderprint.nl
ISBN 978-94-6458-678-7

Towards multicentre application of diffusion MRI in cerebral small vessel disease

Op naar de multi-center toepassing van diffusie MRI in cerebral small vessel disease

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op
woensdag 16 november 2022 des middags te 12.15 uur

door

Bruno M. de Brito Robalo

geboren op 15 mei 1992
te Cape Verde, Kaapverdië

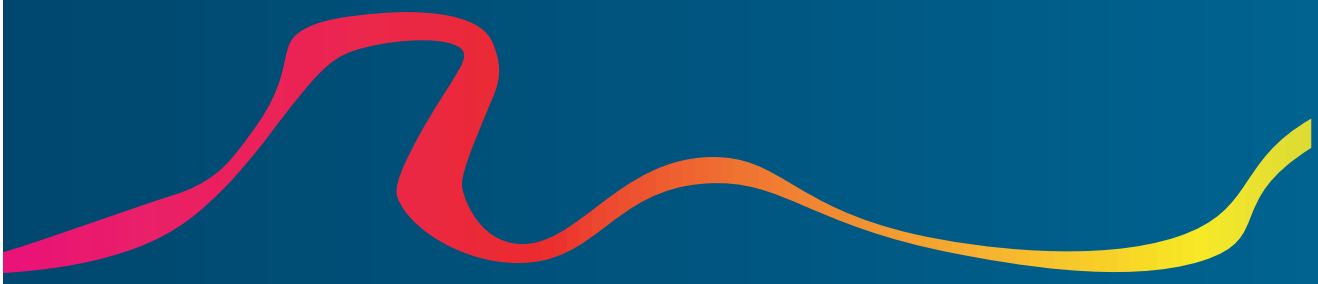
Promotoren	Prof. dr. G.J. Biessels
	Prof. dr. A.L.G. Leemans
Copromotoren	Dr. Y.D. Reijmer
	Dr. A. de Luca
Beoordelingscommissie	Prof. dr. M.J.N.L. Benders (voorzitter)
	Prof. dr. K.P.J. Braun
	Prof. dr. R.M. Dijkhuizen
	Prof. dr. F.E. de Leeuw
	Prof. dr. M. Smits

Let everything happen to you
Beauty and terror
Just keep going
No feeling is final

Rainer Maria Rilke

CONTENTS

Chapter 1



General Introduction

DIFFUSION MRI IN SVD RESEARCH

Cerebral small vessel disease (SVD) is a major cause of cognitive impairment and dementia (Pantoni, 2010). SVD affects the small vessels of the brain and results in characteristic lesions visible with magnetic resonance imaging (MRI), including lacunar infarcts, white matter hyperintensities (WMH), cerebral microbleeds, and brain atrophy. These conventional MRI markers are routinely evaluated in clinical practice but relate poorly with clinical outcome (Smith et al., 2019). Moreover, these markers mainly describe changes in brain tissue that occur at a later stage of the disease, but do not inform about microstructural changes of the underlying tissue. Therefore, it remains difficult to predict clinically relevant symptoms such as cognitive impairment based on imaging, with prognostic value (Pasi & Cordonnier, 2020).

Diffusion MRI (dMRI) on the other hand has the potential to explain SVD-related at the microstructural level and offers more insights into the pathophysiology of the disease (Baykara et al., 2016; Wiegertjes et al., 2019). Diffusion is a physical phenomenon related to the random motion of water molecules, which move differentially according to the organization of underlying tissues. Measures derived from dMRI are sensitive to changes in brain tissues beyond lesions detected with convectional MRI. As an extension of dMRI, fibre tractography allows to reconstruct major white matter pathways and map whole brain networks (Yeh et al., 2021). This framework can be used to study topological changes in brain connectivity and identify tracts-specific damage related to brain functions. Promising research has already shown that a possible avenue through which SVD affects cognition is by disrupting network connections, hampering efficient exchange of information throughout the brain (Morris et al., 2014; Petersen et al., 2020; Reijmer et al., 2015).

Despite its potential, dMRI remains difficult to implement in a clinical setting compared to other imaging markers (Schilling et al., 2019, 2021). An important challenge is the lack of generalization of results across settings. This is mainly due to differences in scanner and acquisition and choice of post-processing steps. Differences in acquisition result in low comparability of dMRI signal across centres. Regarding post-processing, higher order dMRI analyses such as network analysis, require many processing steps. The nature of dMRI network analyses, together with inconsistencies that can accumulate with multiple processing steps can produce networks with many false positive connections. This is an important source of different network architectures between datasets. These inconsistencies due acquisition and post processing affect measurements obtained: i) across different scanners and/or acquisition parameters; ii) in longitudinal studies with long rescan intervals; iii) to compare different groups of subjects, particularly in the presence of pathology. Therefore, it has been difficult to validate SVD dMRI findings in large datasets, which is a requirement for implementation in the clinical setting.

THESIS AIM

Overarching aim: The aim of this thesis is to advance the implementation of dMRI in the context of SVD research, by exploring methods that improve its technical validity. I will focus on two **key objectives**:

- To improve the validity of multicentre dMRI measurements, by removing variability related to acquisition hardware in dMRI across sites with harmonization of the raw diffusion signal.
- To improve the validity of dMRI-based brain network analyses, in the presence of pathology, removing false positive connections by thresholding.

SCANNER-RELATED VARIABILITY OF DMRI

Large multicentre dMRI datasets can help to increase sensitivity to effects of interest and increase internal and external validity of the results of dMRI studies. However, joint analysis of dMRI obtained at different sites and/or acquired at different time points is challenging due to local and/or temporal scanner characteristics that cause in a high inter- and intra-scanner variability (Vollmar et al., 2010). These can be caused by a range of factors such as scanner hardware (Grech-Sollars et al., 2015) (e.g., scanner, transmitter/receiver coils, magnetic field strength, gradient strength, or field inhomogeneities), acquisition parameters (e.g., voxel size, number of gradient directions, echo time), scanner software and reconstruction algorithms (Sotiropoulos et al., 2013) (GRAPPA or SENSE, etc.,).

Once images are acquired, pre-processing and quality control of the data also present sources of variability (Zeilinger et al., 2017). These may include motion correction, eddy current, susceptibility induced distortion correction, processing to reduce partial volume effects, all of which affect the measured diffusion signal intensity, resulting in poor agreement of metrics derived from downstream analysis. Ultimately, it is important to understand that there are many choices to make in acquisition and processing that will affect the raw data, the type and robustness of analyses possible, and the absolute values of quantitative metrics. Therefore, direct comparisons of diffusion metrics between studies or datasets acquired on different scanners cannot be performed without accounting for systematic variability obscuring true effects of interest.

In prospective studies, scanner and acquisition-related variability can be controlled beforehand using standardized acquisitions and scanners from the same manufacturer (Konieczny et al., 2020). Nevertheless, reliability studies have demonstrated that diffusion metrics of subjects acquired using the same scanner and acquisition protocol may still differ significantly across sites. Even in phantoms, scalar dMRI metrics such as mean diffusivity (MD) and fractional anisotropy (FA) have shown more than 7% variability across sites (Palacios et al., 2017;

Teipel et al., 2011). In the human brain, these differences can be non-uniform across different regions, with effect sizes reaching the same order or of magnitude as differences between cases and controls (Kumar et al., 2009). In retrospective studies, where it is no longer possible to adjust scanner and acquisition protocols, differences in dMRI measurements across sites are even more pronounced. Consequently, the majority of retrospective dMRI studies in disease have been single centre with modest sample sizes, using external cohorts only for validation rather than pooled analysis. It is therefore imperative that multicentre dMRI is harmonized prior to joint-analysis.

Several harmonization techniques have been developed and can be divided into two categories. The first category operates on each diffusion metric individually (e.g., FA and MD) by using statistical approaches such as meta-analysis to combine z-scores from all sites or by modelling each site with a covariate during analysis (Palacios et al., 2017). Another method that works on directly diffusion metrics is ComBat (Fortin et al., 2017) (combating batch effects when combining batches), which estimates an additive and a multiplicative site-effect coefficient at each voxel. A limitation of this method is the assumption that site-effect follows a particular parametric distribution, which might not generalize to all scenarios or diffusion metrics. The major downside of both ComBat and of other statistical methods is that each diffusion measure is harmonized separately, which might alter the relation between dMRI metrics that are mathematically linked. Moreover, it is not clear how non-linearities in the signal propagate through pre-processing steps, as well as model fitting procedures (Pinto et al., 2020).

By contrast, the second category of harmonization operates directly on the raw dMRI signal rather than on diffusion metrics (Koppers et al., 2019; Mirzaalian et al., 2015). The advantage of these approaches is that the raw diffusion signal is harmonized in a model-independent manner, allowing any type of subsequent analysis once the data has been harmonized, such as fibre tractography and network analysis. In the studies described in this thesis we focus on the second type of harmonization, and particularly on the rotation invariant spherical harmonics (RISH) method, which maps the dMRI signal from a 'target' site to a 'reference' site. This signal mapping is possible because the voxel-wise dMRI signal intensity can be represented in a spherical harmonics (SH) basis with a given number of parametrization coefficients (Descoteaux et al., 2007). From this representation, RISH features can be calculated and used to scale the dMRI signal from the target site to the reference.

In **chapter 2** we show proof of concept that RISH harmonization can be applied in multicentre analysis of dMRI of elderly subjects with SVD to remove acquisition-related differences in scalar metrics (FA, MD and peak width of skeletonized MD – PSMD) while preserving well-known associations with WMH. In **chapter 5** we further demonstrate that RISH harmonization is beneficial for multicentre network analysis, by improving cross-scanner consistency of dMRI-based brain networks while also improving sensitivity and precision to detect disease effects.

INCONSISTENCIES IN BRAIN NETWORKS ANALYSIS

Fibre tractography is a dMRI technique that allows us to reconstruct white matter pathways and define large scale brain networks. Brain networks are often described as a set of nodes (cortical and subcortical gray matter regions) connected by edges (white matter tracts) (Hagmann et al., 2008). Subsequently, graph theory is used to characterize this network, allowing us to examine topological network changes in psychiatric and neurological disorders (Bullmore & Sporns, 2009). In recent years, this approach has emerged as a useful tool to understand how focal and non-focal degeneration (e.g., Alzheimer's disease and SVD) affect cognitive function.

Despite its widespread use, the process from data acquisition to generation of brain networks is a multi-step procedure with numerous assumptions and uncertainties that can ultimately affect the ability to faithfully represent true axonal connections of the brain. A fundamental problem with brain network analysis is the lack of a ground truth, meaning it is impossible to identify false positive (FP) and false negative (FN) connections in the network (Hein et al., 2016; Maier-Hein et al., 2017). In the context of dMRI-based brain networks, a false positive refers to finding a connection that is either non-existing in reality, or not representative of the underlying anatomy. On the contrary, a false negative represents the failure to reconstruct an actually existing connection.

Several sources of error contribute for the presence of false positives and false negatives in the network. The well-established approach of diffusion tensor imaging (DTI) is unable to resolve complex fibre orientations (Von Dem Hagen & Henkelman, 2002). Streamlines derived with DTI tractography are likely to terminate or choose only one dominant direction in such regions. DTI tractography is therefore prone to many false negatives. In contrast methods such as constrained spherical deconvolution (Tournier et al., 2008) are less prone to false negatives, but can lead to spurious peaks in the fibre orientation distribution, especially in presence of datasets characterized by low angular resolution, low signal to noise ratio or poorly characterized fibre response functions (e.g., due to alterations of the brain microstructure). This will result in more erroneous streamline trajectories and more false positives. When dealing with datasets of patients with presence of pathology, fibre tractography methods, which are developed using high quality datasets of healthy subjects or phantoms, might have a suboptimal performance, potentially resulting different amounts false positives and false negatives (de Reus & van den Heuvel, 2013) and disparities in network density across subjects, as explored in **Chapter 4**. Furthermore, network edges can be weighted by different indices, such as FA or number of streamlines (NOS), which present more sources of variability in network definition.

The most straightforward approach to alleviate the effects of false positives is to employ thresholding to remove 'weak' connections (van Wijk et al., 2010). This can be done at group level where connections with low consistency across groups are removed, or at individual level based on diffusion parameters (e.g., low number of streamlines in the reconstructed tracts).

This procedure is useful on two accounts: firstly, it helps to remove the spurious connections (the false positives discussed above) and secondly by inducing the sparsity it helps minimize the number of connections tested in ensuing analyses, minimizing multiple comparison problems. There are multiple thresholding approaches described in the literature, for example fixed-density, proportional, consensus and consistency thresholding (Garrison et al., 2015). Many network studies have used absolute thresholding (Garrison et al., 2015), where a uniform threshold is applied to remove all connections below a certain edge-weight (e.g., below 5 streamlines). Another popular method is fixed-density thresholding (Rubinov & Sporns, 2010; van den Heuvel et al., 2017), where a relative threshold derived from each individual data is applied to remove the weakest connections, such that an equal network density is achieved across subjects.

Scan-rescan studies in healthy controls suggest that applying thresholds to remove false positives can improve network similarity across datasets and over different time points, not only in terms of graph metrics but also by replicating the same network architecture (e.g., the same set of edges and edge-weights) (Andreotti et al., 2014; Buchanan et al., 2019; Messaritaki et al., 2019; Sarwar et al., 2018; Tsai, 2018; Welton et al., 2015). In **chapter 3**, I will also explore how thresholding methods that improve network reproducibility in controls can also be applied to improve network analysis in longitudinal studies, with patients scanned over longer time frames.

OUTLINE OF THE THESIS

The aim of this thesis is to advance the implementation of dMRI in the context of SVD by exploring methods that improve its technical validity, with a particular focus on harmonization of raw dMRI data and network thresholding. **The two key objectives will be addressed in four chapters:**

- In chapter 2 we assess if RISH harmonization effectively removes acquisition-related differences in multicentre dMRI of subjects with SVD, without affecting disease effects.
- In chapter 3 we examined how thresholding brain networks to a core structure improves network consistency of in SVD patients and healthy controls, while remaining sensitive to group-differences in cross-sectional analysis.
- In chapter 4 we explore the impact of thresholding on longitudinal network analysis, with a focus on consistency of network architecture and sensitivity to disease effects.
- Finally, in chapter 5 we combine RISH harmonization with network thresholding to facilitate large multicentre network analysis with improved consistency of network architecture across scanners and improved sensitivity to detect SVD-related network injury.

REFERENCES

Andreotti, J., Jann, K., Melie-Garcia, L., Dierks, T., Federspiel, A., & Giezendanner, S. (2014). Repeatability Analysis of Global and Local Metrics of Brain Structural Networks. Brain Connectivity, 4(3), 203–220. https://doi.org/10.1089/brain.2013.0202

Baykara, E., Gesierich, B., Adam, R., Tuladhar, A. M., Biesbroek, J. M., Koek, H. L., Ropele, S., Jouvent, E., Chabriat, H., Ertl-Wagner, B., Ewers, M., Schmidt, R., de Leeuw, F.-E., Biessels, G. J., Dichgans, M., & Duering, M. (2016). A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. Annals of Neurology, 80(4), 581–592. https://doi.org/10.1002/ana.24758

Buchanan, C. R., Bastin, M. E., Ritchie, S. J., Liewald, D. C., Madole, J., Drob, M. T., Deary, I. J., & Cox, S. R. (2019). The effect of network thresholding and weighting on structural brain networks in the UK Biobank. 1–29.

Bullmore, E., & Sporns, O. (2009). Complex brain networks : graph theoretical analysis of structural and functional systems. 10(February). https://doi.org/10.1038/nrn2575

de Reus, M. A., & van den Heuvel, M. P. (2013). Estimating false positives and negatives in brain networks. NeuroImage, 70, 402–409. https://doi.org/10.1016/j.neuroimage.2012.12.066

Descoteaux, M., Angelino, E., Fitzgibbons, S., & Deriche, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. Magnetic Resonance in Medicine, 58(3), 497–510. https://doi.org/10.1002/mrm.21277

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. NeuroImage, 161(July), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Garrison, K. A., Scheinost, D., Finn, E. S., Shen, X., & Constable, R. T. (2015). The (in)stability of functional brain network measures across thresholds. NeuroImage, 118, 651–661. https://doi.org/10.1016/j.neuroimage.2015.05.046

Grech-Sollars, M., Hales, P. W., Miyazaki, K., Raschke, F., Rodriguez, D., Wilson, M., Gill, S. K., Banks, T., Saunders, D. E., Clayden, J. D., Gwilliam, M. N., Barrick, T. R., Morgan, P. S., Davies, N. P., Rossiter, J., Auer, D. P., Grundy, R., Leach, M. O., Howe, F. A., … Clark, C. A. (2015). Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. NMR in Biomedicine, 28(4), 468–485. https://doi.org/10.1002/nbm.3269

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Van Wedeen, J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. PLoS Biology, 6(7), 1479–1493. https://doi.org/10.1371/journal.pbio.0060159

Hein, K. H. M., Neher, P., Christophe, J., & Alexandre, M. (2016). Tractography-based connectomes are dominated by false-positive connections. BioRxiv, 1–23. https://doi.org/10.1101/084137

Konieczny, M. J., Dewenter, A., Telgte, A. ter, Gesierich, B., Wiegertjes, K., Finsterwalder, S., Kopczak, A., Hübner, M., Malik, R., Tuladhar, A. M., Marques, J. P., Norris, D. G., Koch, A., Dietrich, O., Ewers, M., Schmidt, R., de Leeuw, F.-E., & Duering, M. (2020). Multi-shell

diffusion MRI models for white matter characterization in cerebral small vessel disease. In Neurology. https://doi.org/10.1212/wnl.0000000000011213

Koppers, S., Bloy, L., Berman, J. I., Tax, C. M. W., Edgar, J. C., & Merhof, D. (2019). Spherical Harmonic Residual Network for Diffusion Signal Harmonization. Mathematics and Visualization, 226249, 173–182. https://doi.org/10.1007/978-3-030-05831-9_14

Kumar, R., Husain, M., Gupta, R. K., Hasan, K. M., Haris, M., Agarwal, A. K., Pandey, C. M., & Narayana, P. A. (2009). Serial changes in the white matter diffusion tensor imaging metrics in moderate traumatic brain injury and correlation with neuro-cognitive function. Journal of Neurotrauma, 26(4), 481–495. https://doi.org/10.1089/neu.2008.0461

Maier-Hein, K. H., Neher, P. F., Houde, J. C., Côté, M. A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F. C., Lin, Y. C., Ji, Q., Reddick, W. E., Glass, J. O., Chen, D. Q., Feng, Y., Gao, C., Wu, Y., Ma, J., Renjie, H., Li, Q., … Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion tractography. Nature Communications, 8(1). https://doi.org/10.1038/s41467-017-01285-x

Messaritaki, E., Dimitriadis, S. I., & Jones, D. K. (2019). Optimization of graph construction can significantly increase the power of structural brain network studies. NeuroImage, 199(May), 495–511. https://doi.org/10.1016/j.neuroimage.2019.05.052

Mirzaalian, H., de Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M., Westin, C. F., Shenton, M. E., & Rathi, Y. (2015). Harmonizing diffusion MRI data across multiple sites and scanners. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9349, 12–19. https://doi.org/10.1007/978-3-319-24553-9_2

Morris, R. G., Lawrence, A. J., Chung, A. W., Barrick, T. R., & Markus, H. S. (2014). Structural network efficiency is associated with cognitive impairment in small-vessel disease. Neurology, 83(4), 304–311. https://doi.org/10.1212/wnl.0000000000000612

Palacios, E. M., Martin, A. J., Boss, M. A., Ezekiel, F., Chang, Y. S., Yuh, E. L., Vassar, M. J., Schnyer, D. M., MacDonald, C. L., Crawford, K. L., Irimia, A., Toga, A. W., & Mukherjee, P. (2017). Toward precision and reproducibility of diffusion tensor imaging: A multicenter diffusion phantom and traveling volunteer study. American Journal of Neuroradiology, 38(3), 537–545. https://doi.org/10.3174/ajnr.A5025

Pantoni, L. (2010). Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. The Lancet Neurology, 9(7), 689–701. https://doi.org/10.1016/S1474-4422(10)70104-6

Pasi, M., & Cordonnier, C. (2020). Clinical Relevance of Cerebral Small Vessel Diseases. Stroke, 51(1), 47–53. https://doi.org/10.1161/STROKEAHA.119.024148

Petersen, M., Frey, B. M., Schlemm, E., Mayer, C., Hanning, U., Engelke, K., Fiehler, J., Borof, K., Jagodzinski, A., Gerloff, C., Thomalla, G., & Cheng, B. (2020). Network Localisation of White Matter Damage in Cerebral Small Vessel Disease. Scientific Reports, 10(1), 1–9. https://doi.org/10.1038/s41598-020-66013-w

Pinto, M. S., Paolella, R., Billiet, T., Van Dyck, P., Guns, P. J., Jeurissen, B., Ribbens, A., den Dekker, A. J., & Sijbers, J. (2020). Harmonization of Brain Diffusion MRI: Concepts and

Methods. Frontiers in Neuroscience, 14(May). https://doi.org/10.3389/fnins.2020.00396

Reijmer, Y. D., Fotiadis, P., Martinez-Ramirez, S., Salat, D. H., Schultz, A., Shoamanesh, A., Ayres, A. M., Vashkevich, A., Rosas, D., Schwab, K., Leemans, A., Biessels, G. J., Rosand, J., Johnson, K. A., Viswanathan, A., Gurol, M. E., & Greenberg, S. M. (2015). Structural network alterations and neurological dysfunction in cerebral amyloid angiopathy. Brain, 138(1), 179–188. https://doi.org/10.1093/brain/awu316

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. NeuroImage, 52(3), 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

Sarwar, T., Ramamohanarao, K., & Zalesky, A. (2018). Mapping connectomes with diffusion MRI: deterministic or probabilistic tractography? Magnetic Resonance in Medicine, April 2018, 1368–1384. https://doi.org/10.1002/mrm.27471

Schilling, K. G., Daducci, A., Maier-Hein, K., Poupon, C., Houde, J. C., Nath, V., Anderson, A. W., Landman, B. A., & Descoteaux, M. (2019). Challenges in diffusion MRI tractography – Lessons learned from international benchmark competitions. Magnetic Resonance Imaging, 57(October 2018), 194–209. https://doi.org/10.1016/j.mri.2018.11.014

Schilling, K. G., Tax, C. M. W., Rheault, F., Hansen, C., Yang, Q., Yeh, F. C., Cai, L., Anderson, A. W., & Landman, B. A. (2021). Fiber tractography bundle segmentation depends on scanner effects, vendor effects, acquisition resolution, diffusion sampling scheme, diffusion sensitization, and bundle segmentation workflow. NeuroImage, 242(July), 118451. https://doi.org/10.1016/j.neuroimage.2021.118451

Smith, E. E., Biessels, G. J., De Guio, F., de Leeuw, F. E., Duchesne, S., Düring, M., Frayne, R., Ikram, M. A., Jouvent, E., MacIntosh, B. J., Thrippleton, M. J., Vernooij, M. W., Adams, H., Backes, W. H., Ballerini, L., Black, S. E., Chen, C., Corriveau, R., DeCarli, C., … Wardlaw, J. M. (2019). Harmonizing brain magnetic resonance imaging methods for vascular contributions to neurodegeneration. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, 11, 191–204. https://doi.org/10.1016/j.dadm.2019.01.002

Sotiropoulos, S. N., Moeller, S., Jbabdi, S., Xu, J., Andersson, J. L., Auerbach, E. J., Yacoub, E., Feinberg, D., Setsompop, K., Wald, L. L., Behrens, T. E. J., Ugurbil, K., & Lenglet, C. (2013). Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: Reducing the noise floor using SENSE. Magnetic Resonance in Medicine, 70(6), 1682–1689. https://doi.org/10.1002/mrm.24623

Teipel, S. J., Reuter, S., Stieltjes, B., Acosta-Cabronero, J., Ernemann, U., Fellgiebel, A., Filippi, M., Frisoni, G., Hentschel, F., Jessen, F., Klöppel, S., Meindl, T., Pouwels, P. J. W., Hauenstein, K. H., & Hampel, H. (2011). Multicenter stability of diffusion tensor imaging measures: A European clinical and physical phantom study. Psychiatry Research - Neuroimaging, 194(3), 363–371. https://doi.org/10.1016/j.pscychresns.2011.05.012

Tournier, J. D., Yeh, C. H., Calamante, F., Cho, K. H., Connelly, A., & Lin, C. P. (2008). Resolving crossing fibres using constrained spherical deconvolution: Validation using diffusion-weighted imaging phantom data. NeuroImage, 42(2), 617–625. https://doi.org/10.1016/j.neuroimage.2008.05.002

Tsai, S. Y. (2018). Reproducibility of structural brain connectivity and network metrics using

probabilistic diffusion tractography. Scientific Reports, 8(1), 1–12. https://doi.org/10.1038/s41598-018-29943-0

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T. T., & Schmidt, R. (2017). Proportional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. NeuroImage, 152(December 2016), 437–449. https://doi.org/10.1016/j.neuroimage.2017.02.005

van Wijk, B. C. M., Stam, C. J., & Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. PLoS ONE, 5(10). https://doi.org/10.1371/journal.pone.0013701

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P., & Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. NeuroImage, 51(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Von Dem Hagen, E. A. H., & Henkelman, R. M. (2002). Orientational diffusion reflects fiber structure within a voxel. Magnetic Resonance in Medicine, 48(3), 454–459. https://doi.org/10.1002/mrm.10250

Welton, T., Kent, D. A., Auer, D. P., & Dineen, R. A. (2015). Reproducibility of Graph-Theoretic Brain Network Metrics: A Systematic Review. Brain Connectivity, 5(4), 193–202. https://doi.org/10.1089/brain.2014.0313

Wiegertjes, K., Ter Telgte, A., Oliveira, P. B., Van Leijsen, E. M. C., Bergkamp, M. I., Van Uden, I. W. M., Ghafoorian, M., Van Der Holst, H. M., Norris, D. G., Platel, B., Klijn, C. J. M., Tuladhar, A. M., & De Leeuw, F. E. (2019). The role of small diffusion-weighted imaging lesions in cerebral small vessel disease. Neurology, 93(17), E1627–E1634. https://doi.org/10.1212/WNL.0000000000008364

Yeh, C. H., Jones, D. K., Liang, X., Descoteaux, M., & Connelly, A. (2021). Mapping Structural Connectivity Using Diffusion MRI: Challenges and Opportunities. Journal of Magnetic Resonance Imaging, 53(6), 1666–1682. https://doi.org/10.1002/jmri.27188

Zeilinger, M. G., Lell, M., Baltzer, P. A. T., Dörfler, A., Uder, M., & Dietzel, M. (2017). Impact of post-processing methods on apparent diffusion coefficient values. European Radiology, 27(3), 946–955. https://doi.org/10.1007/s00330-016-4403-6

Chapter 2

Diffusion MRI harmonization enables joint-analysis of multicentre data of patients with cerebral small vessel disease

Bruno M. de Brito Robalo
Geert Jan Biessels
Christopher Chen
Anna Dewenter
Marco Duering
Saima Hilal
Huiberdina L. Koek
Anna Kopczak
Bonnie Yin Ka Lam
Alexander Leemans
Vincent Mok
Laurien P. Onkenhout
Hilde van den Brink
Alberto de Luca

Neuroimage: Clinical 2021 Nov, 32: 213-1582
DOI: 10.1016/j.nicl.2021.102886

ABSTRACT

Objectives: Acquisition-related differences in diffusion magnetic resonance imaging (dMRI) hamper pooling of multicentre data to achieve large sample sizes. A promising solution is to harmonize the raw diffusion signal using rotation invariant spherical harmonic (RISH) features, but this has not been tested in elderly subjects. Here we aimed to establish if RISH harmonization effectively removes acquisition-related differences in multicentre dMRI of elderly subjects with cerebral small vessel disease (SVD), while preserving sensitivity to disease effects.

Methods: Five cohorts of patients with SVD (N = 397) and elderly controls (N = 175) with 3 Tesla MRI on different systems were included. First, to establish effectiveness of harmonization, the RISH method was trained with data of 13 to 15 age and sex-matched controls from each site. Fractional anisotropy (FA) and mean diffusivity (MD) were compared in matched controls between sites using tract-based spatial statistics (TBSS) and voxel-wise analysis, before and after harmonization. Second, to assess sensitivity to disease effects, we examined whether the contrast (effect sizes of FA, MD and peak width of skeletonized MD - PSMD) between patients and controls within each site remained unaffected by harmonization. Finally, we evaluated the association between white matter hyperintensity (WMH) burden, FA, MD and PSMD using linear regression analyses both within individual cohorts as well as with pooled scans from multiple sites, before and after harmonization.

Results: Before harmonization, significant differences in FA and MD were observed between matched controls of different sites ($p < 0.05$). After harmonization these site-differences were removed. Within each site, RISH harmonization did not alter the effect sizes of FA, MD and PSMD between patients and controls (relative change in Cohen's $d = 4\ \%$) nor the strength of association with WMH volume (relative change in $R^2 = 2.8\ \%$). After harmonization, patient data of all sites could be aggregated in a single analysis to infer the association between WMH volume and FA ($R^2 = 0.62$), MD ($R^2 = 0.64$), and PSMD ($R^2 = 0.60$).

Conclusions: We showed that RISH harmonization effectively removes acquisition-related differences in dMRI of elderly subjects while preserving sensitivity to SVD-related effects. This study provides proof of concept for future multicentre SVD studies with pooled datasets.

1 INTRODUCTION

Combining data from multicentre studies is becoming increasingly important in neuroimaging, with the aim to increase statistical power and provide outcomes that are more generalizable than those obtained at single-centre level (Lorca-Puls et al., 2018, Van Horn and Toga, 2014). However, joint analysis of multicentre magnetic resonance imaging (MRI) can be challenging if inter-site variability due to acquisition-related inconsistencies is not considered (Vollmar et al., 2010).

Inter-site variability is particularly problematic in diffusion MRI (dMRI) and can be caused by a range of factors, including scanner hardware (e.g., scanner manufacturer, magnetic field strength, gradient strength, field inhomogeneities), software, acquisition parameters (e.g., voxel size, number of gradient directions, echo time) (Helmer et al., 2016). All these factors may affect the measured diffusion signal intensity and metrics derived from the data. In prospective multicentre studies, this variability can be controlled using standardized acquisitions and scanners from the same manufacturer (Konieczny et al., 2020). However, when retrospectively combining data form different cohorts, differences in acquisition can be substantial. Even with phantoms, dMRI metrics have shown more than 7% variability across sites (Palacios et al., 2017, Teipel et al., 2011, Timmermans et al., 2019). In the human brain, this variability is even more pronounced and non-uniform across tissues with acquisition-related differences reaching the same order of magnitude as case-control differences (e.g., traumatic brain injury vs. controls, Kumar et al., 2009). In such scenarios, if multicentre data were naively pooled into a single analysis, true biological effects would likely be masked by acquisition-related differences. It is therefore crucial that multicentre dMRI is harmonized prior to joint-analysis (Tax et al., 2019).

Two main kinds of retrospective dMRI harmonization techniques have been developed to date. The first category operates on each diffusion metric individually (e.g., fractional anisotropy – FA, and mean diffusivity – MD) by using statistical approaches such as meta-analysis, or by modelling the difference between sites with covariates during analysis (e.g., ComBat, Fortin et al., 2017). By contrast, the second category of harmonization operates directly on the raw dMRI data rather than on each diffusion metric (Karayumak et al., 2019, Koppers et al., 2019, Mirzaalian et al., 2015). This type of harmonization is more general since the raw diffusion signal is harmonized in a model-independent manner, theoretically allowing any type of subsequent analysis. In this study, we focus specifically on the second type of harmonization with the rotation invariant spherical harmonics (RISH) methods (Mirzaalian et al., 2015).

The core idea of the RISH method is to map the dMRI signal from a 'target' site to a 'reference' site, using groups of healthy subjects matched for factors such as age, sex, etc. This signal mapping is possible because the dMRI signal intensity can be represented in a spherical harmonics (SH) basis with a given number of parametrization coefficients (Tournier et al.,

2004). From this representation, RISH features can be extracted and scaled to harmonize the dMRI signal between two sites (Karayumak et al., 2019). Applications of RISH harmonization have been presented using synthetic data and with data of healthy young subjects, with recent work showing that acquisition-related differences are removed while preserving age- and sex-related effects (Karayumak et al., 2019). Recently, this method has also been applied to harmonize a large dataset of patients with Schizophrenia and investigate changes across the lifespan (Cetin-Karayumak et al., 2020). However, the applicability of the RISH method to older individuals exhibiting brain atrophy or to patients with (diffuse) white matter lesions remain unclear.

In this work, we evaluate the RISH harmonization framework in the context of a retrospective multicentre analysis of individuals with brain lesions due to cerebral small vessel disease (SVD). SVD is a leading cause of cognitive impairment and dementia, and it is often investigated with dMRI (Baykara et al., 2016, Lyoubi-Idrissi et al., 2017, Wiegertjes et al., 2019). The patterns of diffusion change in SVD are well documented using the diffusion tensor model, with patients typically exhibiting widespread increase of MD, peak width of skeletonized MD (PSMD) and decrease of FA, often related to white matter hyperintensity (WMH) burden (Tuladhar et al., 2015, Van Norden et al., 2012). Hence, this patient population is well-suited to investigate the efficacy of harmonization methods. Using scans from five SVD cohorts acquired on different systems (Philips Healthcare, NL, and Siemens Healthineers, DE) and with different protocols, we aimed to establish if application of the RISH method removes acquisition-related differences in dMRI of elderly subjects, while preserving the sensitivity to disease effects in SVD. Finally, we show proof of concept of how multicentre harmonized data can be pooled to perform robust inference of the relation between WMH burden and dMRI metrics.

2 METHODS

2.1 Datasets and inclusion criteria

For this retrospective analysis, we obtained scans from five cohorts including healthy elderly subjects and patients with SVD. These cohorts differed in study design and inclusion criteria (described below), comprising four samples with sporadic SVD and one sample with genetically defined SVD (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy, CADASIL). For the present study, we used a harmonized definition for patients and controls. Patients with sporadic SVD had symptomatic SVD defined as a) history of stroke, with a corresponding small subcortical infarct visible on MRI or b) cognitive complaints and presence of WMH burden on MRI (Fazekas score ≥ 2, Fazekas et al., 1987). The presence of CADASIL was confirmed by molecular genetic testing of the NOTCH3 gene

or ultrastructural analysis of a skin biopsy (detection of pathognomonic granular osmiophilic material, Wollenweber et al., 2015). Patients were excluded if they had other major neurological or psychiatric conditions (e.g., multiple sclerosis, epilepsy, Parkinson's disease). Controls had no history of stroke or cognitive complaints for which they sought medical advice, and their MRI did not show signs of lacunes or WMH with Fazekas score ≥ 2. All subjects had a structural MRI (T1-weighted) and a dMRI scan. Characteristics of the study samples included in this study (397 patients and 175 controls) are provided in Tables 1. All studies included in this analysis were approved by the ethics committees of the respective institutions and all participants provided written informed consent.

2.1.1 Utrecht: Patients (n = 171) were selected from the Utrecht participants of the Parelsnoer study memory clinic cohort (Aalten et al., 2014). Age-matched controls (n = 53) were recruited from a community-based cohort (Reijmer et al., 2013). All MRI scans for both cohorts were acquired on the same 3 Tesla Philips scanner (Achieva, Philips, Best, the Netherlands). T1-weighted scans for both cohorts were acquired with the following parameters: voxel size: 1×1×1 mm^3, echo time (TE): 4.5 ms and repetition time (TR): 7.9 ms. dMRI data were obtained with a voxel size: 2.5 × 2.5 × 2.5 mm^3, TR/TE 6638/73 ms, 45 diffusion gradients directions with a b-value of 1200 s/mm2, and 1 b = 0 s/mm^2 averaged 3 times. Fluid-attenuated inversion recovery (FLAIR) images were obtained with TR/TE/inversion time (TI): 11000/125/2800 ms, voxel size: 1 × 1 × 3 mm^3.

2.1.2 Hong Kong: Patients (n = 20) and controls (n = 20) were selected from a community-based cohort, the Chinese University of Hong Kong–Risk Index for Subclinical brain lesions in Hong Kong (CU-RISK) (Lam et al., 2019). MRI scans were acquired on a 3 Tesla Philips scanner (Achieva, Philips, Best, the Netherlands). T1-weighted images were obtained with TR/TE: 7.49/3.46 ms, voxel size: 0.60×1.04×1.04 mm^3 and dMRI had a TR/TE: 8944/60 ms, voxel size: 1×1×2 mm^3; 32 diffusion gradient directions with b-value 1000 s/mm^2 and 1 b = 0 s/mm2. FLAIR images were acquired with TR/TE/TI: 8000/328.6/2400 ms, voxel size: 0.55 × 0.44 × 0.44 mm^3.

2.1.3 Munich: Patients (n = 72) with CADASIL and controls (n = 34) were selected from the prospective VASCAMY (Vascular and Amyloid Predictors of Neurodegeneration and Cognitive Decline in Nondemented Subjects (Baykara et al., 2016) study. All MRI scans were acquired on a 3 Tesla Magnetom Verio scanner (Siemens Healthineers, Erlangen, Germany). T1-weighted scans were obtained using TR/TE: 2500/4.73 ms, voxel size: 1×1×1 mm^3 and DWIs were performed with a voxel size: 2 × 2 × 2 mm^3, TR/TE: 12700/81 ms, 30 diffusion gradient directions with a b-value of 1000 s/mm2, and 1 b = 0 s/mm2. FLAIR images were obtained with TR/TE/TI: 5000/395/1800 ms, voxel size: 1 × 1 × 1 mm^3.

2.1.4 Utrecht2: A second dataset from the UMC Utrecht consisted of patients (n = 34) and controls (n = 18) from an ongoing prospective observational cohort study Zoom@SVDs

(van den Brink et al., 2021). MRI scans were acquired using the same scanner system and acquisition parameters as the Utrecht1 dataset. However, since multiple scanner software and hardware (coil) updates occurred between the two studies, scans from the Zoom@SVDs study are treated as a separate site.

2.1.5 Singapore: Patients (n = 100) and controls (n = 50) were selected from a community-based cohort, the Epidemiology of Dementia in Singapore (EDIS) study (Hilal et al., 2013). All MRI scans were performed on a 3 Tesla Siemens Magnetom Trio Tim scanner (Siemens Healthineers, Erlangen, Germany). T1-weighted scans were obtained with TR/TE: 2300/1.9 ms, voxel size: 1×1×1 mm^3 and dMRI were acquired with a TR/TE: 6800/85 ms, voxel size: 3.1×3.1×3 mm^3; 61 diffusion gradient directions with b-value 1150 s/mm^2 and 7 b = 0 s/mm2. FLAIR images were obtained with TR/TE/inversion time (TI): 9000/82/2500 ms, voxel size: 1 × 1 × 3 mm^3.

2.2 MRI data pre-processing

All datasets were pre-processed using ExploreDTI version 4.8.6 (Leemans et al., 2009) and the Functional Magnetic Resonance Imaging of the Brain (FMRIB) software library (FSL, v6.0.1). Images were corrected for signal drift (Vos et al., 2016), eddy currents, subject motion with rotation of the B-matrix (Leemans and Jones, 2009), and susceptibility induced distortions (Veraart et al., 2013). dMRI data were nonlinearly registered to the T1 and resampled to an isotropic resolution of 2 × 2 × 2 mm3 and brain masks were generated using brain extraction (BET) tool from FSL.

All images were visually inspected to exclude the presence of major artifacts and misregistration. Since the dMRI data were acquired with different b-values in the different cohorts, we adjusted for differences in b-values as part of the harmonization pipeline. We estimated the signal of all dMRI data to a common b-value (b = 1000 s/mm2) using a linear scaling of the signal decay (S/S0) in the logarithmic domain (Jensen et al., 2005, Steven et al., 2014). This signal decay has been validated in phantoms and healthy controls and shown to be robust for datasets with closely spaced b-values centred around b = 1000 s/mm2 where the diffusion signal is not heavily weighted towards non-Gaussian effects (Magin et al., 2019). This approach was also utilized by (Karayumak et al., 2019) for the RISH harmonization method. In our case, b-value scaling was applied to Utrecht1, Utrecht2 and Singapore, which had original b-values of 1200 s/mm2 and 1150 s/mm2.

WMH volumes were segmented from the FLAIR images using an automated pipeline (coroflo) and registered the MNI152 template (Kuijf et al., 2019). All volumes were normalized to the percentage of intracranial volume (ICV) of the MNI brain.

Table 1: Demographics and imaging parameters of the study samples.

	Utrecht 1		Hong Kong		Munich		Utrecht 2		Singapore	
	Controls (N = 46)	Patients (N = 170)	Controls (N = 20)	Patients (N = 20)	Controls (N = 28)	Patients (N = 54)	Controls (N = 18)	Patients (N = 26)	Controls (N = 54)	Patients (N = 359)
Demographics										
Age, years	71.2 ± 4.8	74.9 ± 9.0	69.2 ± 3.4	74.1 ± 3.3	71.8 ± 7.4	56.0 ± 7.1	62.4 ± 6.9	64.8 ± 6.8	66.6 ± 4.8	71.3 ± 6.0
Male sex (%)	28 (61)	99 (59)	10 (50)	10 (50)	16 (57)	20 (37)	10 (55)	18 (69)	31 (52)	166 (46)
MRI markers										
WMH volume (% ICV)	0.3 [0.1, 0.6]	1.2 [0.5, 2.7]	0.1 [0.05, 0.1]	0.6 [0.3, 1.1]	0.2 [0.1 0.7]	6.2 [3.5, 10]	0.03 [0.02, 0.08]	0.7 [0.3, 1.3]	0.04 [0.02, 0.09]	0.8 [0.08, 1.16]
WMH (Fazekas)	0 [0, 1]	2 [1, 2]	0.5 [0, 1]	2 [2, 3]	0 [0, 1]	3 [2, 3]	1 [1,1]	2 [2,3]	1 [1,1]	2 [1,3]
Lacunes (present)	0 (0)	69 (36)	0(0)	-	0 (0)	40 (74)	0(0)	12 (46)	0 (0)	118 (33)
Imaging parameters										
Scanner	3T Philips Achieva		3T Philips Achieva		3T Siemens Verio		3T Philips Achieva		3T Siemens Magnetom Trio, Tim	
Software version	R3.1		MR Release 5.1		syngo MR B19		MR Release 5.6.0		syngo MR B19	
Voxel size (mm^3)	2.5 × 2.5 × 2.5		1 × 1 × 2		2 × 2 × 2		2.5 × 2.5 × 2.5		3.1 × 3.1 × 3.0	
b-value (s/mm^2)	1200		1000		1000		1200		1150	
# Of directions	45		32		30		45		61	

Data presented as mean ± SD, number (percentages) or median [interquartile range]; SVD = small vessel disease; CADASIL=cerebral autosomal-dominant arteriopathy with subcortical infarcts and leukoencephalopathy; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment; WMH = white matter hyperintensity; ICV = intracranial volume

2.3 Harmonization with RISH features

Harmonization of dMRI with rotation invariant spherical harmonics (RISH) features was first proposed by Mirzaalian et al. 2015, with recent improvements allowing harmonization of scans with different acquisition parameters (Karayumak et al., 2019), which is the case in our study. This type of harmonization is based on the fact that dMRI signal along unique gradient directions can be represented with a basis of spherical harmonics (SH). From this representation, RISH features that describe different aspects of the signal can be calculated. RISH features can be viewed as the total energy at a specific angular frequency (order) in the SH space. The core assumption of this method is that two groups of healthy subjects matched for age, sex, lesion burden, etc., are expected to have similar diffusion profiles on a group level and thus none of the RISH features should be statistically different between sites. Under this assumption, eventual group differences observed in diffusion measurements such as FA and MD are attributed to scanner-related inconsistencies, as previously shown in healthy controls (Ning et al., 2020).

To ensure that the average of RISH features captures site properties on a group level and not characteristics of individuals, a minimum number of training controls (15–20) is required from each site (Karayumak et al., 2019). Subsequently, a scaling is determined between the average of RISH features such that scanner-related differences are removed between sites. This mapping is linear in the SH domain, but non-linear in the original diffusion signal domain. We provide a detailed description of RISH harmonization in the supplementary information, and further theory can also be found in the original method papers (Karayumak et al., 2019; Mirzaalian et al., 2015)

In short, RISH harmonization pipeline consists of two parts: 1) learning inter-site differences in the form of scale maps between RISH features of the reference and target site (Figure 1, part 1); and 2) applying the learned scale maps to harmonize all dMRI datasets of the target site (Figure 1, part 2). The learning part is performed using age- and sex-matched controls as training data. From the dMRI signal, the RISH features are calculated and registered to a common spatial template generated from training subjects with ANTs (Avants et al., 2010). In the template space, the expected values of RISH features are defined as the sample mean over the number of training subjects and voxel-wise scale maps of RISH features are estimated between the target and reference site. Next, in the application part, the scale maps are warped to the subject space and used to harmonize the SH coefficients of the target site. Finally,

the harmonized diffusion signal can be reconstructed.

2.4 Experimental design and analysis

2.4.1 Effectiveness of RISH harmonization

Here we assess if acquisition-related differences in diffusion metrics between sites can be removed by RISH harmonization. The first step was to select Training Controls from every cohort that were as similar as possible between sites to minimize sources of variability other than scanner. The Utrecht1 cohort was used as a reference site because the age range of the controls allowed matching with all other sites. This was done on a site-by-site basis, generating four sets of Training Controls with participants from every site matched for age and sex to participants from Utrecht1 (demographics in supplementary Table S1). Tract-based-spatial-statistics (TBSS, Smith et al., 2006) and voxel-based analysis were used to compare the FA and MD between Training Controls of the reference and target sites, before and after harmonization. For the TBSS pipeline, FA and MD were estimated using the diffusion tensor model (*dtfit* from FSL). Next, FA maps were aligned to the MNI152 template and a white matter skeleton representing the centres of major bundles was computed. Subsequently, FA and MD of the skeleton were compared between reference and target sites in a voxel-wise fashion using t-tests with threshold-free cluster enhancement (5000 permutations). This comparison was also extended to the whole brain to ensure that acquisition-related differences in grey matter regions and other structures are also removed.

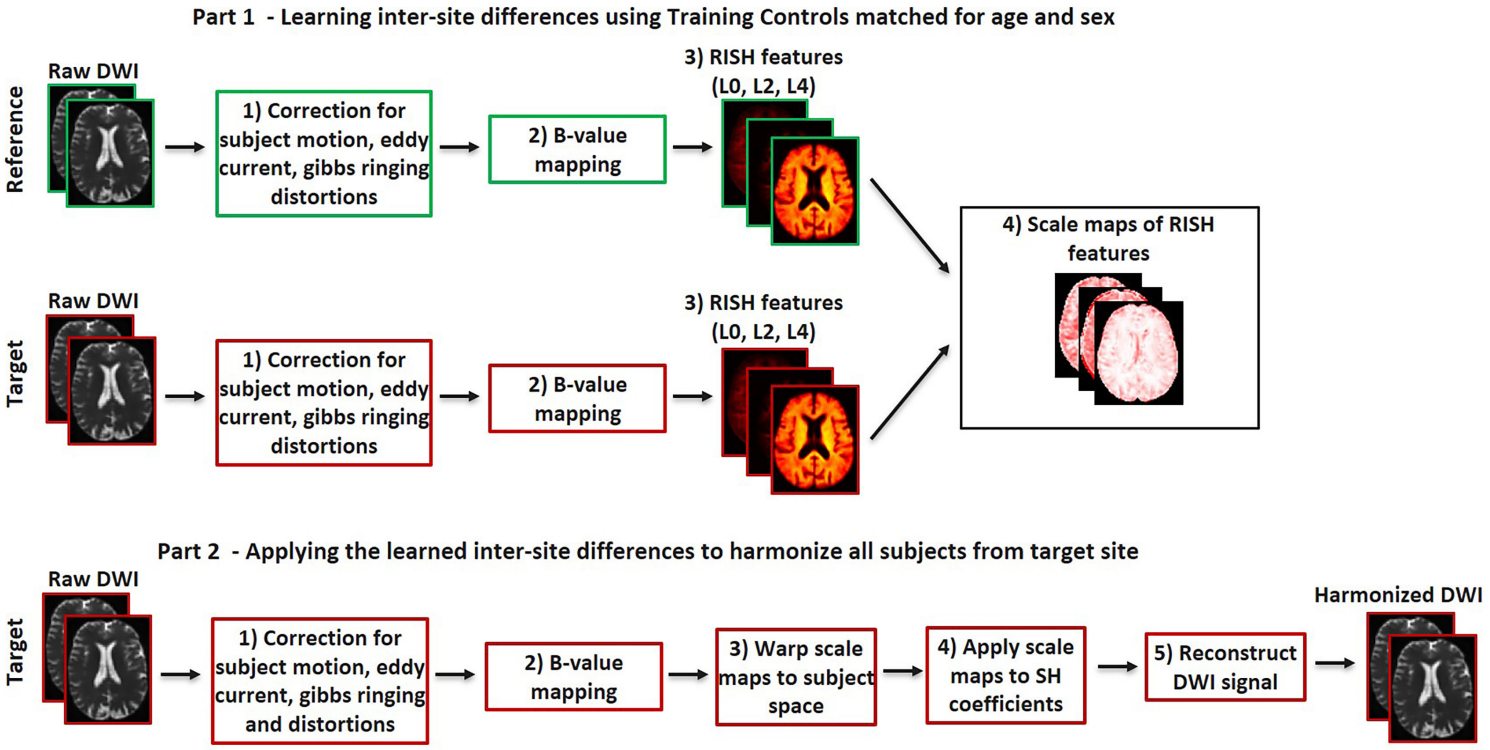


Figure 1: Harmonization steps using RISH features. Part 1) All scans are pre-processed to correct for artefacts, followed by b-value mapping to a common b-value of 1000 s/mm2. Voxel-wised scale maps are computed using a set of Training Controls matched for age and sex between the reference and the target site. Part 2) The scale maps are then applied to harmonize the remaining scans of the target site.

We also evaluated the generalizability of effectiveness of harmonization beyond the Training Controls. This was done by creating a group of Validation Controls with data from Utrecht1 (n = 15), Munich (n = 15) and Singapore (n = 15), since those sites had a sufficient number of controls outside the Training Controls to generate separate sets of matched groups (demographics in supplementary Table S2). Similar to the analysis with training controls, TBSS and voxel-based analysis were used to compare validation controls between each target site and the reference, before and after harmonization. Furthermore, one-way ANOVA was performed to compare the average FA and MD of the TBSS skeleton across these three sites.

2.4.2 Sensitivity to disease effects in SVD

For this objective we included all patients from all sites (demographics in Table 1). We made a selection of controls that included both training and validation controls, while ensuring that the average age was similar across sites (demographics in supplementary Table S3). We matched controls for age across sites in order to have a common reference to calculate effect sizes against patients. Thus, after harmonization controls are expected to have similar diffusion measures while the contrast with their respective patient groups should not be affected. To assess sensitivity to disease effects, general linear model (GLM) adjusted for age and sex was performed to examine the contrast (effect sizes of FA, MD and PSMD from the TBSS skeleton) between patients and controls within each site, before and after harmonization.

In patients, we also explored the sensitivity to disease effects by relating WMH volume with FA, MD and PSMD, before and after harmonization. Linear regression adjusted for age and sex was performed. Since the WMH volumes was non-normally distributed, a Box-Cox transformation was applied (Box & Cox, 1964). We determined the R2 and standardized regression coefficients (β), before and after harmonization.

2.4.3 Proof of concept of data pooling

We evaluated if disease effects were similar if patients were compared to Internal Controls or to a pooled set of matched controls derived from external centres only. GLM adjusted for age and sex was performed to compare patients from Utrecht1 and from Munich versus External Controls pooled from other sites. We compared effect sizes of FA, MD, PSMD obtained with External Controls to the original effect sizes with Internal Controls.

Finally, we demonstrated proof of concept of data pooling by relating WMH volume

with FA, MD and PSDM on pooled data of patients from multiple sites and compared the fit of the curve before and after harmonization. Similar to the analysis within sites, linear regression was performed.

3 RESULTS

3.1 Effectiveness of RISH harmonization

Figure 2 shows RISH features of order 0, 2 and 4 obtained from the Training Controls and corresponding scale maps, before harmonization. We observed widespread differences in RISH features between the reference and target sites that were dependent on the tissue type. Small differences were observed in regions with prevalently single fibre populations (e.g., corpus callosum) for all orders of RISH features, whereas more peripheral white matter and grey matter regions showed bigger differences across sites (see scale maps). Furthermore, data from Hong Kong, Munich and Singapore sites showed larger differences from the reference (Utrecht1) than Utrecht2.

Before harmonization, significant differences in FA and MD were found between the reference and each target site across the entire brain (Figure 3), especially for Hong Kong Munich and Singapore ($p<0.05$). After harmonization, all significant differences in the white matter skeleton between the target sites and the reference were removed. When analysing the whole brain, FA differences were still seen for the Hong Kong site ($p < 0.05$), mainly in subcortical grey matter and near tissue interfaces with cerebrospinal fluid, probably due to misregistration (Figure 3, top right panel). A map of effect sizes further clarifies that differences (positive or negative Cohen's d) are removed after harmonization (i.e., effect sizes become closer to zero, Supplementary Figure S1).

Regarding Validation Controls (Figure 4), voxel-wise differences in FA and MD in the white matter skeleton and across the whole brain were removed after harmonization, with exception of minor differences at tissue interfaces, probably due to misregistration. When comparing the average FA and MD of the skeleton significant differences in FA were found across sites before harmonization ($F(1,43) = 18.2$, $p < 0.001$, Figure 5A). All differences in FA were removed after harmonization. The average MD of the skeleton did not differ across sites before or after harmonization (Figure 5B).

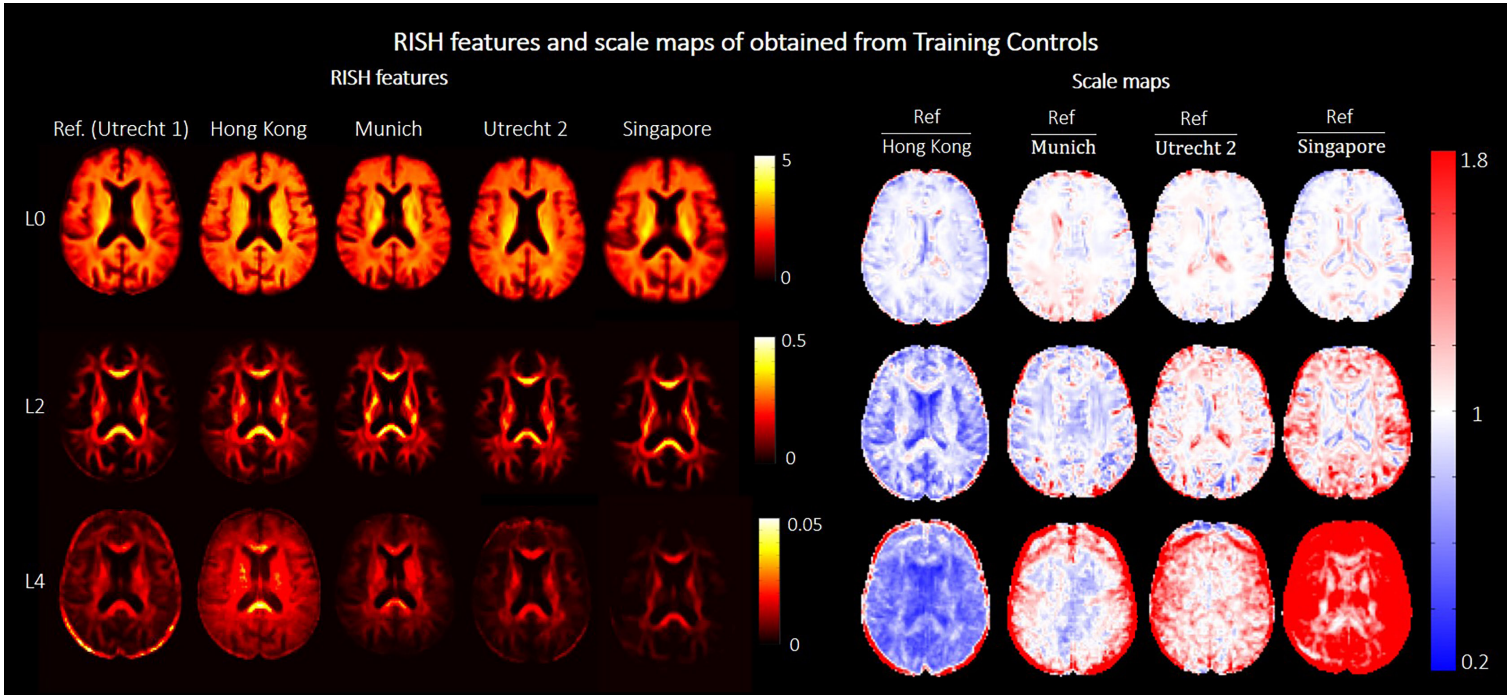


Figure 2: Left (top to bottom): RISH features of order 0, 2 and 4 calculated with Training Controls of each site. Different columns correspond to the different sites. Right (top to bottom): Scale maps for each RISH feature obtained by scaling each target site to the reference. Different columns correspond to different target sites. The red-blue colormap indicates the scaling factor between the two sites.

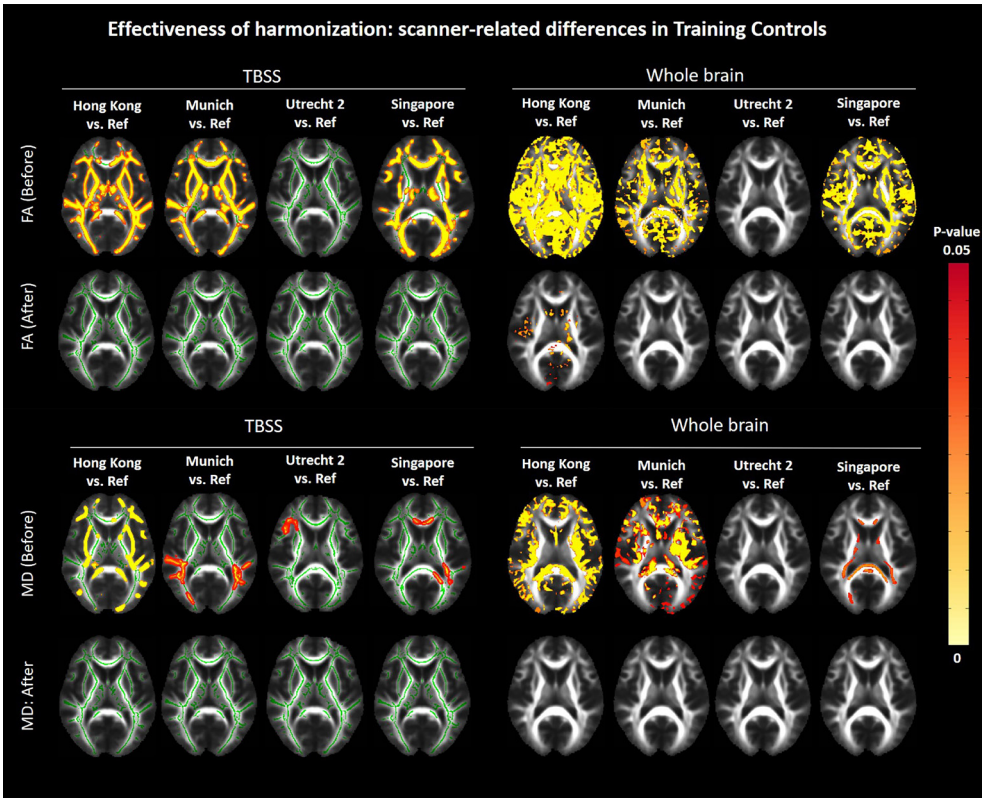


Figure 3: Results of the TBSS (left) and whole brain voxel-wise analysis (right) comparing FA (top) and MD (bottom) between Training Controls of each target site and the reference, before and after harmonization. The yellow-red colormap shows voxels where statistical differences were observed after multiple comparison corrections (p-value < 0.05). Corresponding maps of effect sizes are shown in Supplementary material (Part 3, Figure S1).

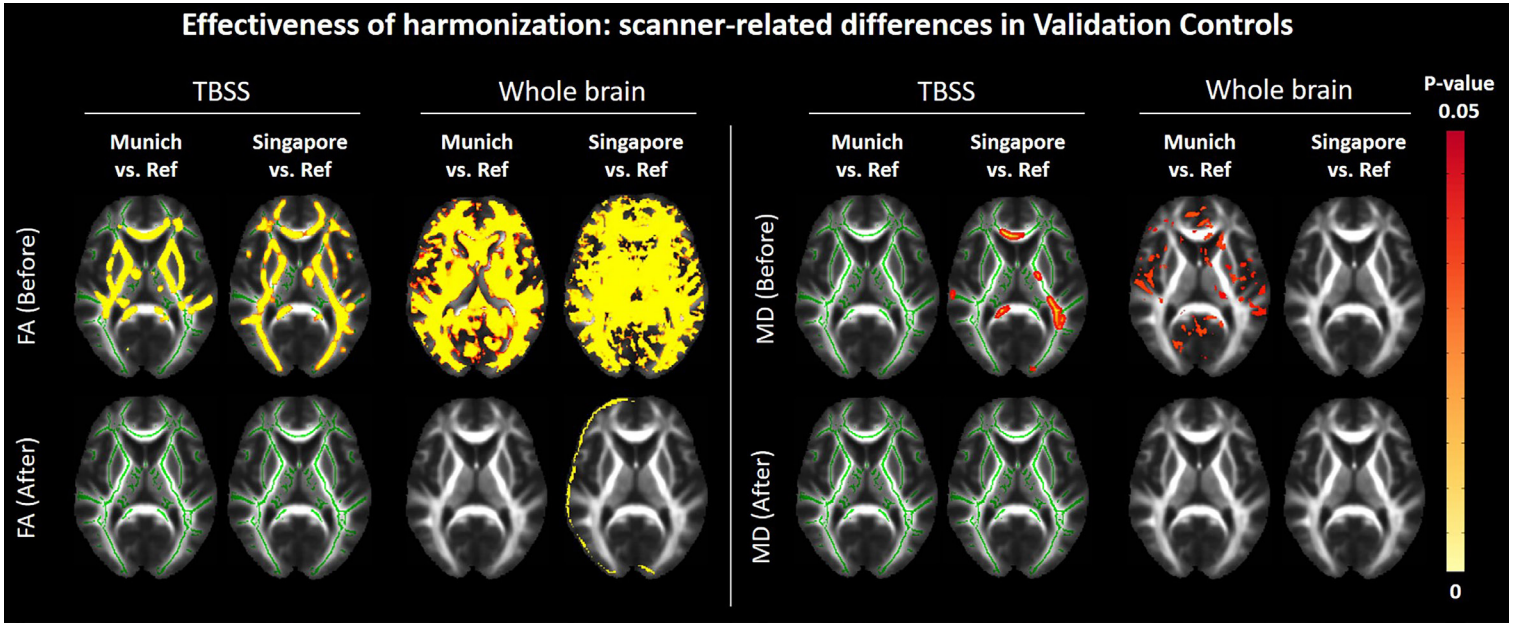


Figure 4: Results of the TBSS and whole brain voxel-wise analysis comparing FA (left) and MD (right) between Validation Controls of each target site and the reference, before and after harmonization (demographics in supplementary table S2). The yellow-red colormap shows voxels where statistical differences were observed after multiple comparison corrections (p-value < 0.05).

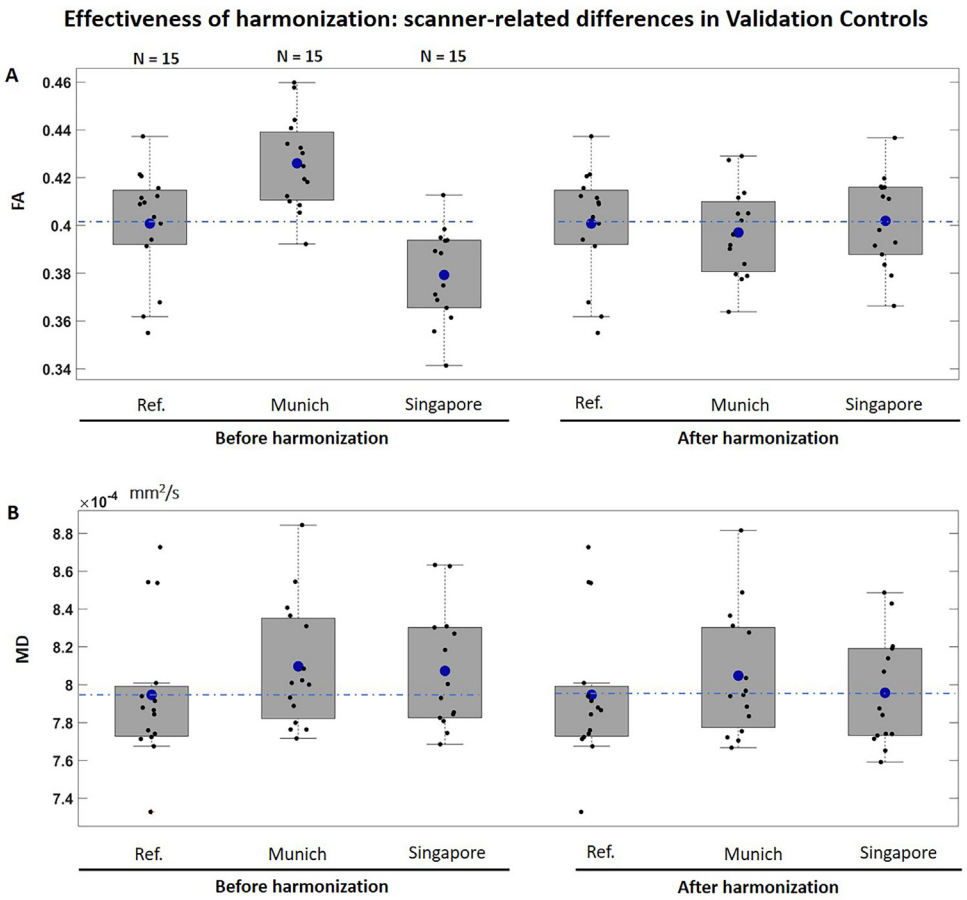


Figure 5: Boxplots of average FA (panel A) and MD (panel B) of the white matte skeleton compared between Validation Controls, before and after harmonization (demographics in supplementary table S2). The dashed blue line represents the mean of the reference site. The blue marker within each boxplot indicates the mean of the corresponding group, which were compared between sites using a one-way ANOVA. Top: FA before harmonization, $F(1,43) = 18.2$, $p<0.001$; FA after harmonization, $F(3,43) = 0.2$, $p = 0.8$. Bottom: MD before harmonization: $F(3,43) = 0.83$, $p = 0.06$; MD after harmonization: $F(3,43) = 2.6$, $p = 0.4$.

3.2 Sensitivity to disease effects in SVD

Figure 6 depicts differences in dMRI metrics between patients and controls within each site, with quantitative values shown in Table 2. Before harmonization, patients had a significantly lower FA (Figure 6A), higher MD (Figure 6B) and higher PSMD (Figure 6C) than controls in all sites except Hong Kong: FA (d = -0.96 to -2.07, p<0.001); MD (d = 1.02 to 1.99, p<0.001); PSMD (d = 0.93 to 1.71, p<0.001). After harmonization, all effect sizes were preserved, regardless of if they were small (0.2), medium (0.5) or large (0.8). On average, the relative change in effect size from pre- to post-harmonization was 3.9 % (Table 2). Voxel-wise analysis of one of the target sizes (Munich) shows that regional differences between patients and controls are preserved after harmonization (Supplementary Figure S2).

Before harmonization, WMH volume was significantly associated with all dMRI metrics in all sites, FA (R^2 = 0.37 to 0.68, p < 0.001); MD (R^2 = 0.57 to 0.70, p<0.001); PSMD (R^2 = 0.49 to 0.76; p<0.001), except Hong Kong where associations MD were not significant (Figure 7). After harmonization, all associations were preserved regardless of the strength, with R^2, standardized β coefficients being marginally affected. The relative change in R^2 after harmonization was 2.8%.

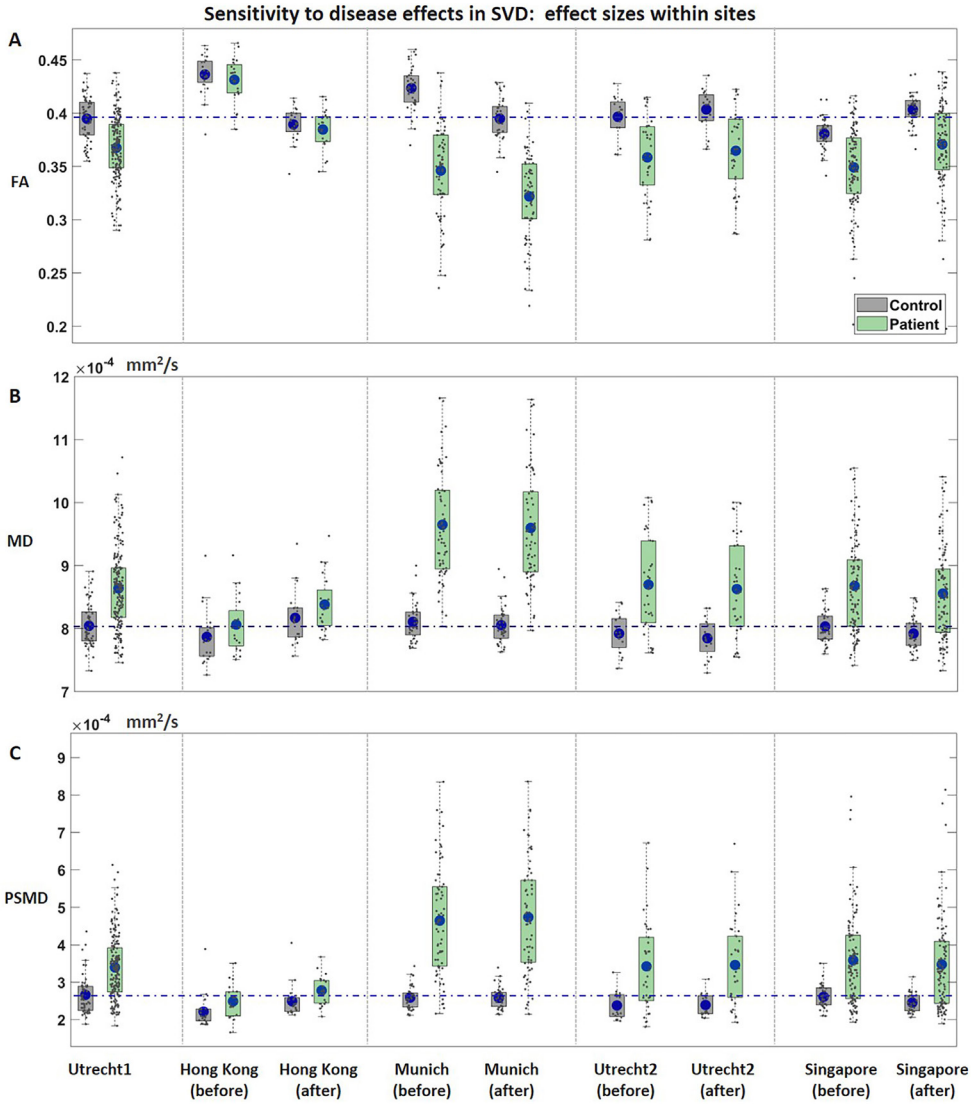


Figure 6: Average FA (panel A), MD (panel B) and PSMD (panel C) of the white matter skeleton compared between patients (green) and controls (gray) within each site (demographics for selected controls in supplementary table S3). Results are displayed for the reference site and for each target site before and after harmonization. Corresponding p-values and effect sizes are displayed in Table 2. The dashed line indicates the mean value of controls of the reference site.

Table 2: Effect sizes of FA, MD and PSMD between patients and controls within sites, before and after harmonization.

		Utrecht1	Hong Kong before	Hong Kong after	% Change	Munich before	Munich after	% Change	Utrecht2 before	Utrecht 2 after	% Change	Singapore before	Singapore after	% Change
FA	P-value	1.40×10^{-11}	.43	.39		3.17×10^{-21}	2.30×10^{-21}		2.20×10^{-5}	2.20×10^{-5}		3.30×10^{-10}	9.00×10^{-15}	
	Effect size (Cohen's d)	-.96	-.25	-.27	8	-2.07	-2.09	1	-1.16	-1.16	0	-1.19	-1.17	2
	CI	(-1.29, -.64)	(-.87, .37)	(-.88, .35)		(-2.57, -1.57)	(-2.59, -1.59)		(-1.79, -.51)	(-1.79, -.52)		(-1.54, -.83)	(-1.53, -.81)	
MD	P-value	6.29×10^{-10}	0.17	0.13		9.57×10^{-21}	6.29×10^{-21}		2.84×10^{-4}	2.62×10^{-4}		2.40×10^{-14}	5.80×10^{-14}	
	Effect size (Cohen's d)	1.02	.44	.48	9	1.99	2	1	1.18	1.19	1	1.14	1.11	3
	CI	(.63, 1.34)	(-.19, 1.06)	(-.15, 1.11)		(1.49, 2.49)	(1.49, 2.48)		(.54, 1.82)	(.55, 1.83)		(.80, 1.50)	(.76, 1.50)	
PSMD	P-value	3.10×10^{-12}	.06	.04		3.6×10^{-17}	5.97×10^{-18}		3.2×10^{-6}	9×10^{-6}		4.10×10^{-12}	2.40×10^{-12}	
	Effect size (Cohen's d)	.93	.62	.69	11	1.71	1.78	4	1.03	1.1	7	-.99	-.99	0
	CI	(.62, 1.25)	(-.02, 1.25)	(.05, 1.32)		(1.23, 2.18)	(1.29, 2.25)		(.39, 1.65)	(.46, 1.73)		(-1.34, -.64)	(-1.35, -.65)	

Demographics for selected controls in Supplementary Table S3.

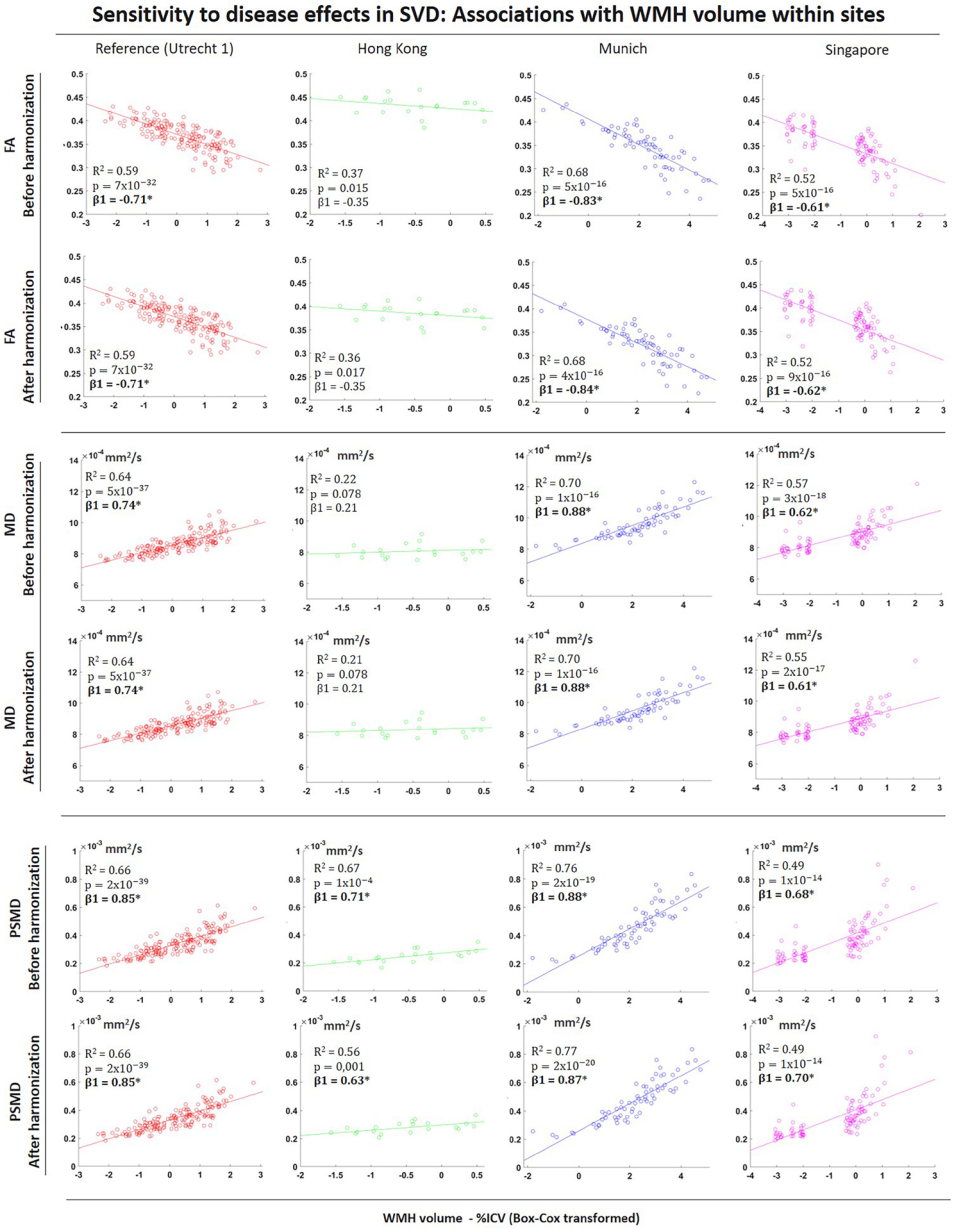


Figure 7: Scatter plots of associations between WMH volume, FA (top), MD (middle) and PSMD (bottom) within sites, before and after harmonization. MD and PSMD values are given in mm2/s. *$p < 0.05$ for β coefficients.

3.3 Proof of concept of data pooling

When comparing Utrecht1 patients to External Controls before harmonization, differences in FA were close to twice as large as the original effect size obtained with internal controls from Utrecht1 (d = -1.87, compared to -0.96, Figure 8A). After harmonization, the effect size of FA between Utrecht1 patients and External Controls was more comparable to the original effect sizes (Cohen's d = -1.1, compared to -0.96). Results were similar when we performed the same analysis using patients from Munich (Figure 8B)): effect sizes between patients and the External Controls were more similar to the original effect size after harmonizing the data (original effect size: d = -2.07; effect size with Pooled Controls before harmonization d = -1.67, after harmonization d = -2.1). For MD and PSMD, the effect sizes between patients and External Controls were similar to the original effects for both Utrecht1 and Munich, even before harmonization (Figure 8 C-F).

Regarding associations between WMH and FA on the pooled data before harmonization, FA values from different sites were clustered in separate clouds (Figure 9A). This non-harmonized data still described a significant association between WMH volume and FA but with weaker correlations than some individual sites due to the clustering effect ($R^2 = 0.33$; $p = 2\times10^{-31}$). After harmonization, data points were more aligned around the fitted curve, with the measurements behaving as a single centre data (Figure 9B). This resulted in stronger associations between WMH volume and FA ($R^2 = 0.62$; $p = 2\times10^{-75}$). For MD (Figure 9 C-D) and PSMD (Figure 9 E-F), the clustering of points was less prominent, but associations with WMH volume also became stronger after harmonization. Before harmonization, MD ($R^2 = 0.61$; $p = 7\times10^{-74}$); PSMD ($R^2 = 0.56$; $p = 6\times10^{-64}$); after harmonization, (MD: $R^2 = 0.64$; $p = 7\times10^{-89}$); PSMD ($R^2 = 0.60$; $p = 5\times10^{-71}$).

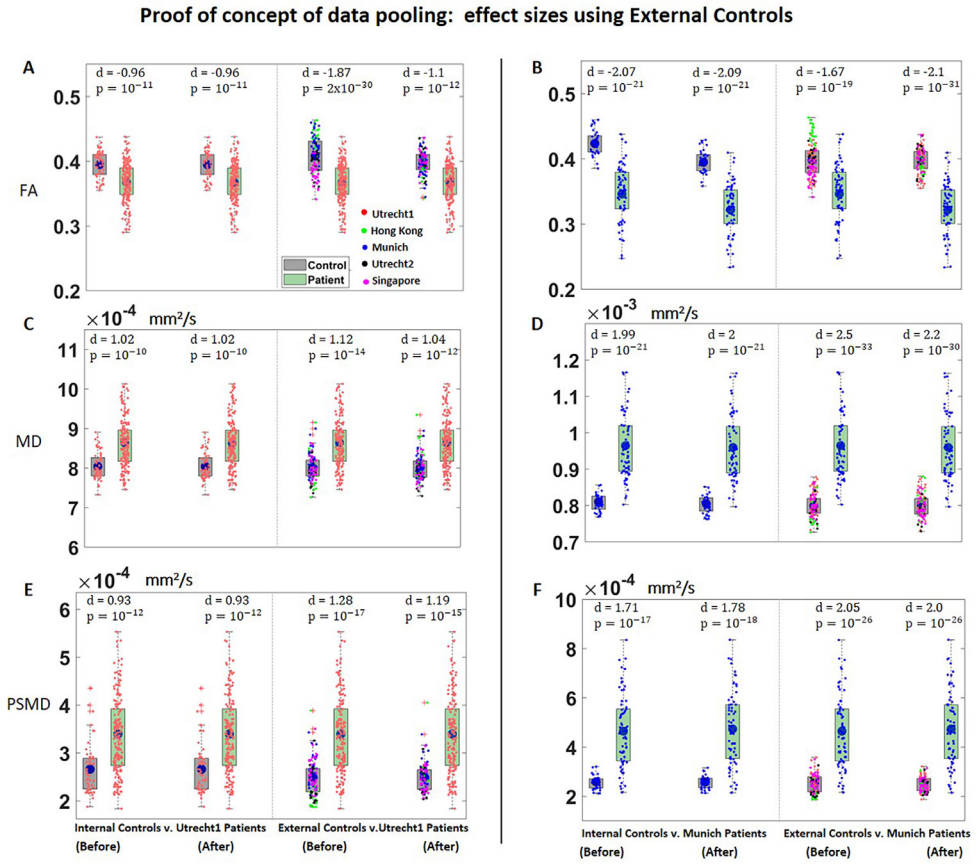


Figure 8: Boxplots of FA (A-B), MD (C-D) and PSMD (E-F) of the white matter skeleton compared between patients and controls. (Left): Comparisons between Utrecht1 patients and Internal Controls and between Utrecht1 patients and External Controls, before and after harmonization. (Right): Comparisons between Munich patients and Internal Controls, and between Munich patients and External Controls, before and after harmonization (demographics for selected controls in supplementary table S3). Data from each site is colour coded as follows: red = Utrecht1; green = Hong Kong; blue = Munich; black = Utrecht2; magenta: Singapore.

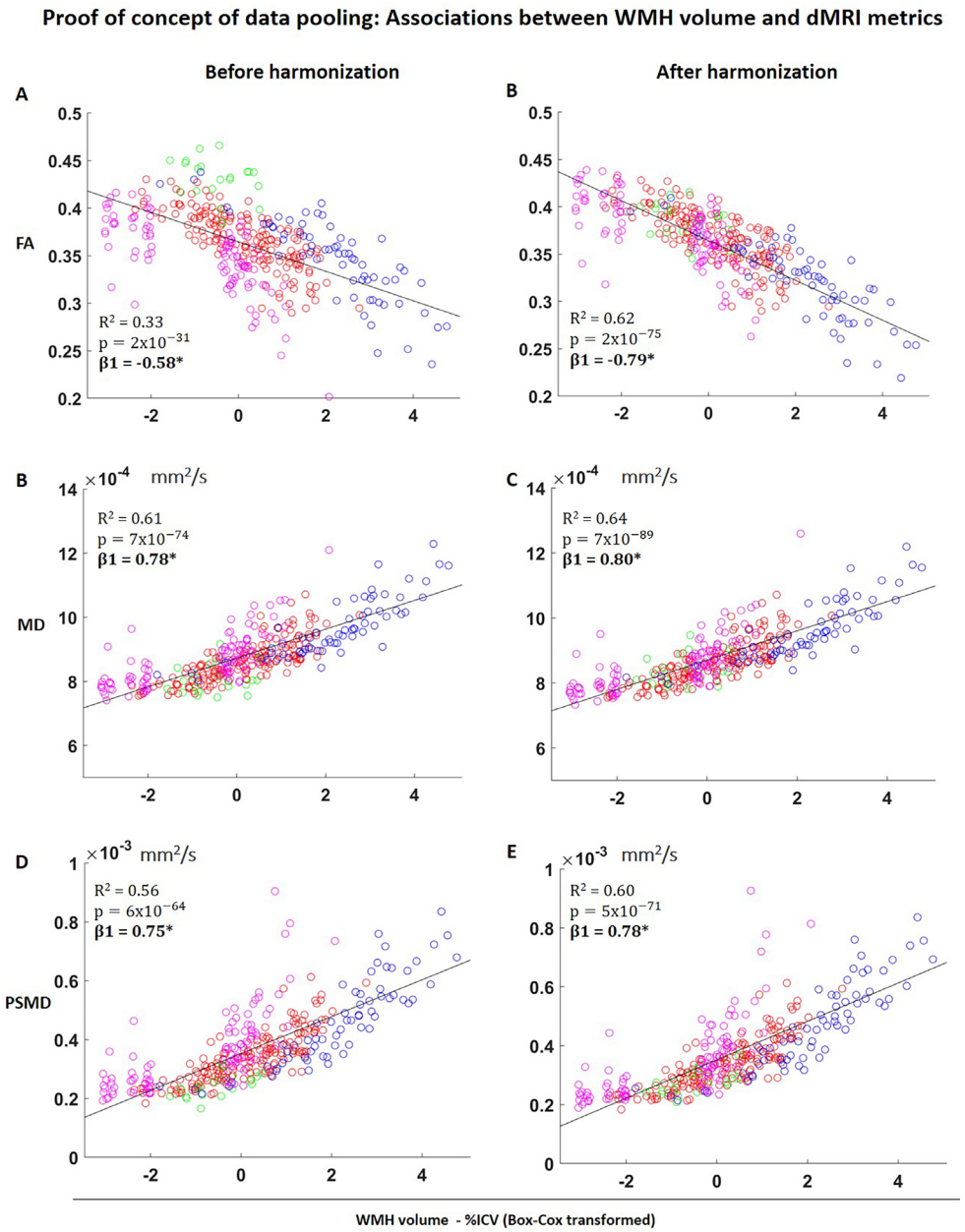


Figure 9: Scatter plots of associations between WMH volume, FA (top), MD (middle) and PSMD (bottom) on the pooled data, before and after harmonization. MD and PSMD values are given in mm2/s. *p < 0.05 for β coefficients. Data from each site is colour coded as follows: red = Utrecht1; green = Hong Kong; blue = Munich; magenta: Singapore.

4 DISCUSSION

We investigated the applicability of RISH harmonization to remove acquisition-related differences in multicentre dMRI of elderly subjects with SVD while preserving disease-related effects. Before harmonization, we observed significant differences in FA and MD across sites, which were removed after harmonization, both in the Training Controls and in Validation Controls not involved in the training step. Importantly, effect sizes of FA, MD and PSMD for group differences between patients and controls as well as for associations with WMH volume within each site were preserved after harmonization. The harmonized controls could be effectively considered as a single-site dataset. The pooled data of patients covered a wide range of WMH burden, allowing to demonstrate a strong relation between WMH volume and dMRI metrics.

The RISH method has been previously implemented using scans of healthy young subjects without apparent brain lesions for the training step (Cetin-Karayumak et al., 2020, Mirzaalian et al., 2015). Here, we have evaluated whether peculiar characteristics of the elderly brain, such as presence of lesions as WMH brain atrophy and larger ventricles, which are present to some extent even in control subjects, could affect the computation of the scale maps of RISH features. In our study, the selection criteria for controls were based on low burden of SVD. Thus, the training controls were minimally affected by WMH and had relatively similar brain volumes. The largest differences in RISH features were observed in grey matter and peripheral white matter areas, where partial volume effects might play a role on the diffusion profile (Vos et al., 2011, Vos et al., 2012). After harmonization differences in FA and MD between the target sites and the reference were removed across the brain, except for Hong Kong where minor differences in MD still persisted in deep grey matter structures and at tissue interfaces with cerebrospinal fluid. This is likely due to residual inaccuracies in image registration or differences in WMH burden and brain volumes in Training controls, given these were not explicitly matched for these markers (Supplementary Figure S7). Accordingly, we suggest that when dealing with data of elderly subjects it might be beneficial to match Training Controls not only for age and sex (Hsu et al., 2008), but also in terms of WMH lesion distribution, brain volumes or other demographics that contribute for variation in diffusion (e.g., handedness, race, etc., Büchel et al., 2004), although this might be challenging to achieve in practice in most studies. This is particularly important when the inclusion criteria for controls are not based on MRI markers of SVD but rather on variables such as being cognitively healthy.

Another important consideration for studies implementing the RISH method is that imaging artefacts specific of one site or few subjects (e.g., ghosting, incomplete fat saturation) might be learned as part of the harmonization features and propagate into the harmonized dataset. In our study, an example of these artefacts can be seen as rings due to incomplete fat suppression on the L4 scale maps shown in Figure 2. Nevertheless, their impact on the harmonized data was deemed minimal and it did not significantly affect any subsequent result. The RISH method does not assume that two groups of healthy older subjects are completely identical, but as shown by our work, if groups are matched for major factors, differences in RISH features can be learned

on a group-level, without major influence of individual properties. This is further supported by our results on the generalizability of RISH harmonization with Validation Controls. We demonstrated not only that dMRI metrics of subjects not involved in the training step are harmonized, but also the transitivity of harmonization, e.g., that the independent harmonization of two target sites to the reference also implies harmonization between target sites.

Next, we demonstrated that RISH harmonization does not affect the sensitivity of dMRI to effects of SVD. Well-known differences in FA, MD and PSDM between patients and controls observed within each site before harmonization were preserved (Baykara et al., 2016, Wiegertjes et al., 2019). Effect sizes were unaffected after harmonization (relative change = 3.9 %.), regardless of whether the magnitude was small, medium or large. We believe such small change in effect size can be deemed negligible and is likely caused by registration and interpolation inaccuracies (Karayumak et al., 2019). We also repeated the same analysis with the statistical harmonization method ComBat (Fortin et al., 2017) for comparison. RISH harmonization outperformed ComBat in this dataset, which was not able to preserve effect sizes within all sites (see supplementary Information, Part 6). Recent work has shown that application of the RISH method does not alter the relation between dMRI metrics and biological effects such as age-related changes (Karayumak et al., 2019; 2020). Here, we extended such finding by showing that RISH harmonization also preserved the relation between WMH volume and dMRI metrics, a well-established relation in this kind of patients (Van Leijsen et al., 2017), regardless of the strength or the sample size available within each site to test such associations.

To date, most studies of SVD with dMRI have been single site based, and the inclusion of cohorts from other sites, which can differ substantially in terms of acquisition, has been limited to external validation only. Our results with External Controls indeed show that multicentre data without standardized acquisition across centres cannot be simultaneously analysed, as their integration before harmonization would result in biased effect sizes when comparing patients and controls. Effect sizes obtained with External Controls before harmonization were biased up to 1 standard deviation, which is on the same order of magnitude as typical differences between patients and controls. Conversely, after harmonization External Controls behaved as single-site dataset that could be used as reference for patients from all sites with minimal bias in effect sizes. An important implication of this result is that harmonization can potentially address data obsolescence. Diffusion scans are routinely acquired at single-site level for the purpose of testing specific hypotheses and discarded afterwards as hardware updates are implemented or the acquisition protocols are adjusted. Being able to account for such differences might allow to include previously acquired data across multiple studies, thus valorising previous investments and reducing the burden of scanning new controls in prospective studies.

Another potential benefit of harmonization is in longitudinal studies where upgrades in scanner systems complicate the comparison of data at different time points (Takao et al., 2012). Still, for prospective multi-centre studies and especially clinical trials, which typically lack a healthy control group needed for post-hoc data harmonization, standardization of the acquisition should

still have a high priority.

After establishing that RISH harmonization allows to integrate data from different cohorts, we demonstrated associations between WMH volume and dMRI metrics on the pooled dataset of patients, which resulted into improved statistical power due to the larger sample size. Since the cohorts included in this study had different disease burdens, the pooled data covered a larger spectrum of WMH volumes, allowing to test associations with more confidence than what the individual sites would allow. Since WMH volume already has strong correlations with dMRI metrics (Bendlin et al., 2010), even the non-harmonized pooled data resulted in significant associations. However, R2 were lower than some individual sites, showcasing again the risk of performing inferences using non-harmonized data. After harmonization, data points of different sites behaved as a single-centre data and more aligned with the fitted curve, resulting in an increase of R2. The impact of harmonization is likely to be even more important in the study of other correlates of SVD with more subtle effect sizes than WMH (e.g., relation between dMRI metrics and cognition, Du et al., 2020) or when assessing disease progression over time (van Leijsen et al., 2019).

5 LIMITATIONS AND FUTURE DIRECTIONS

Despite its advantages, RISH harmonization does not come without limitations. To minimize differences in diffusion weighting across sites, we used a linear scaling to map the diffusion signal to a b-value of 1000 s2/mm, which is only applicable to a limited range of b-values. For prospective multicentre studies, it is crucial to minimize differences in acquisition parameters such as b-values, but for retrospective studies with already acquired data, harmonization of scans with largely different b-values needs further investigation. As clinical protocols continuously improve and more complex sequences are implemented (e.g., multi-shell data), future work should also investigate whether RISH harmonization is suitable for such advanced dMRI applications in SVD (De Luca et al., 2018, Konieczny et al., 2020, Rydhög et al., 2017). Moreover, since we focused our analyses on the dMRI metrics most commonly associated with SVD (FA, MD, PSMD), further analyses are required in order to generalize our conclusions to other metrics obtained from higher level analysis such as fibre tractography (De Luca et al., 2020) and network theory in SVD (Reijmer et al., 2015).

6 CONCLUSIONS

Despite the limitations, our study is the first to prove the feasibility of RISH harmonization of multicentre dMRI scans in the context of SVD. We showed that harmonizing the raw dMRI signal is effective in removing acquisition-related differences, while preserving the sensitivity to disease effects. This ultimately allowed us to directly pool scans acquired at different sites into a single analysis and increase the power of dMRI inferences. Our work paves the way not only for the validation of dMRI markers of SVD in large scale multicentre studies, but also for studies that aim to answer new research questions where statistical power is critical, such as

untangling underlying aetiologies in SVD populations (e.g., free-water imaging, Finsterwalder et al., 2020). When translated to even larger scales, harmonization could significantly improve the sensitivity and specificity for studies attempting to identify tract specific damage and network connections related to cognitive dysfunction (Konieczny et al., 2020).

ACKNOWLEDGEMENTS

We acknowledge and thank all patients and controls for study participation. We thank members of the Utrecht Vascular Cognitive Impairment (VCI) Study group. We also thank Mathias Hübner for technical assistance and Angelika Doerr as study nurse for her help in the VASCAMY study in Munich.

SUPPLEMENTARY INFORMATION

Part 1 - Tables

Table S1: Demographic characteristics Training Controls used to learn scanner differences and compute voxel maps of RISH features between the reference and target sites.

	Hong Kong vs Reference		Munich vs Reference		Utrecht2 vs Reference		Singapore vs Reference		Munich vs Reference (retraining)	
	Hong Kong (N = 15)	Reference (N = 15)	Munich (N = 15)	Reference (N = 15)	Utrecht2 (N = 13)	Reference (N = 13)	Singapore (N = 15)	Reference (N = 15)	Munich (N = 15)	Reference (N = 15)
Demographics										
Age, years	69.0 ± 3.2	68.7 ± 3.4	70.6 ±3.5	70.8 ±4.4	65.0 ±3.0	65.5 ±0.5	70.6 ±2.8	70.8±4.4	69.4 ±5.1	69.4 ±4.1
Male sex (%)	47	47	53	53	61	53	53	53	47	47
Total brain volume (%ICV)	78.1 ± 9.1	74.4 ± 5.8	76.9 ± 3.2	73.9 ± 3.8	81.1 ± 8.2	79.4 ± 6.0	78.1 ± 9.1	74.1 ± 9.4	79.0 ± 4.1	74.8 ± 5.5

Each target site (Hong Kong, Munich, Utrecht2) was matched to the reference site (Utrecht1) for age and sex. In the matched groups, no significant differences were observed for both covariates ($p > 0.05$). Munich data was training twice using a second selection of Training Controls, to ensure that harmonization is not affected by sampling bias (results in section 2 of the supplementary information). ICV – Intracranial volume.

Table S2: Demographic characteristics of Validation Controls.

	Reference (N=15)	Munich (N=15)	Singapore (N=15)
Demographics			
Age, years	70.8 ± 3.4	70.9 ± 5.7	68.6 ± 5.6
Male sex (%)	67	47	60

For each site, controls that were not involved in the training step were selected as Validation Controls, to test generalizability of harmonization. This was done for Utrecht1, Munich and Singapore, since those sites had a sufficient number of controls to generate separate sets of matched groups.

Table S3: Demographic characteristics of controls used to test effect sizes vs. patients.

	Utrecht1 (N = 44)	Hong Kong (N = 20)	Munich (N = 30)	Singapore (N = 34)	Utrecht2 (N = 12)
Demographics					
Age, years	68.8 ± 3.6	69.2 ± 3.4	69.3 ± 4.3	68.3 ± 4.9	68.4 ± 3.4
Male sex (%)	54	50	40	59	66

This selection of controls includes both Training and Validation Controls. Age-matching of controls across sites was applied to reduce between site differences in controls. We made no attempt to also age-match the patients across sites, as this was not required for the intended analyses.

Part 2 - Detailed descriptions of RISH harmonization

2.1 Diffusion MRI signal and RISH features

The use of rotation invariant spherical harmonics (RISH) for harmonization was first proposed by Mirzaalian et al. (2015), with recent improvements allowing harmonization of datasets with different acquisition parameters (Karayumak et al, 2019), which is the case in our study. This harmonization method is based on the fact that dMRI signal $S = [S_1 ... S_G]^T$ along unique gradient directions (G) can be accurately represented in a basis of spherical harmonics (SH) with coefficients C_{ij} given by: $S \approx \sum\sum C_{ij} Y_{ij}$, where Y_{ij} are SH basis functions of order i and phase j. From this representation, several RISH features $\mathcal{F}$ per harmonic order are computed as follows (Mirzaalian et al. 2015):

$$\mathcal{F} = \|C_0\|^2 \|C_2\|^2 \dots \|C_8\|^2 \text{ where } \|C_i\|^2 = \sum_{j=1}^{2i+1} (C_{ij})^2 \qquad (1)$$

RISH features can be viewed as the total energy of the signal in a particular frequency band (order) in the SH space and capture different aspects of the signal. Given two groups of healthy subjects matched for age and sex, they are expected to have similar diffusion profiles and thus the RISH features should not be statistically different on a group level. Therefore, differences found in diffusion measurements are mainly attributed to scanner-related inconsistencies. During the harmonization procedure, the goal is to determine appropriate scaling factors for RISH features between the reference and target sites. This mapping is linear in the SH domain, but non-linear in the original diffusion signal domain.

2.2 Mapping voxel-wise RISH features in training controls

We estimated voxel-wise linear mapping between each target site and the reference site using training controls, as depicted in Part 1 of Figure 1. We computed 3 RISH features of order 0, 2 and 4 which capture different microstructural aspects of the signal. We did not calculate higher order RISH features since in our dataset they primarily capture noise. Next, RISH feature maps were registered to create a multi-modal template using ANTs algorithm (Avants et al., 2010). In the template space, the expected values of the voxel-wise RISH features are defined as the sample mean over the number of training subjects:

$$\mathbb{E}_i^s \approx \sum_{n=1}^{N_t} [\|C_i^n\|^2] / N_s \qquad (2)$$

Where C^n_i is the RISH feature of order i for the nth subject and N_s is the number of training subject for site s. From here voxel-wise scale maps between RISH features of target site (TAR) and reference site (REF) are computed in the template space using:

$$\phi\,(REF, TAR) \approx \sqrt{\frac{\mathbb{E}_i^{REF}}{\mathbb{E}_i^{TAR} + \varepsilon}} \qquad (3)$$

Where ε is a small nonzero constant.

2.3 Applying scale maps to harmonize all DWIs

The scale maps estimated from the training data were applied to all other scans of the target site. First the scale maps are nonlinearly transformed to the native subject space. This transformation is obtained by registering the RISH features of each subject to the template space and then inverting the transformation. The dMRI signal can now be recalculated by scaling the SH coefficients at each voxel in the subject space with:

$$\hat{C}_{ij} = \phi\,(REF, TAR)\hat{C}_{ij} \qquad (4)$$

Where $\hat{C}_{ij}$ are the scaled SH coefficients. Finally, the harmonized diffusion signal in the SH domain is transformed back to the signal intensity domain as follows:

$$\hat{S} = \sum\sum \hat{C}_{ij}\, Y_{ij} \qquad (5)$$

.

Part 3 - Effectiveness of harmonization in training controls: effect size maps

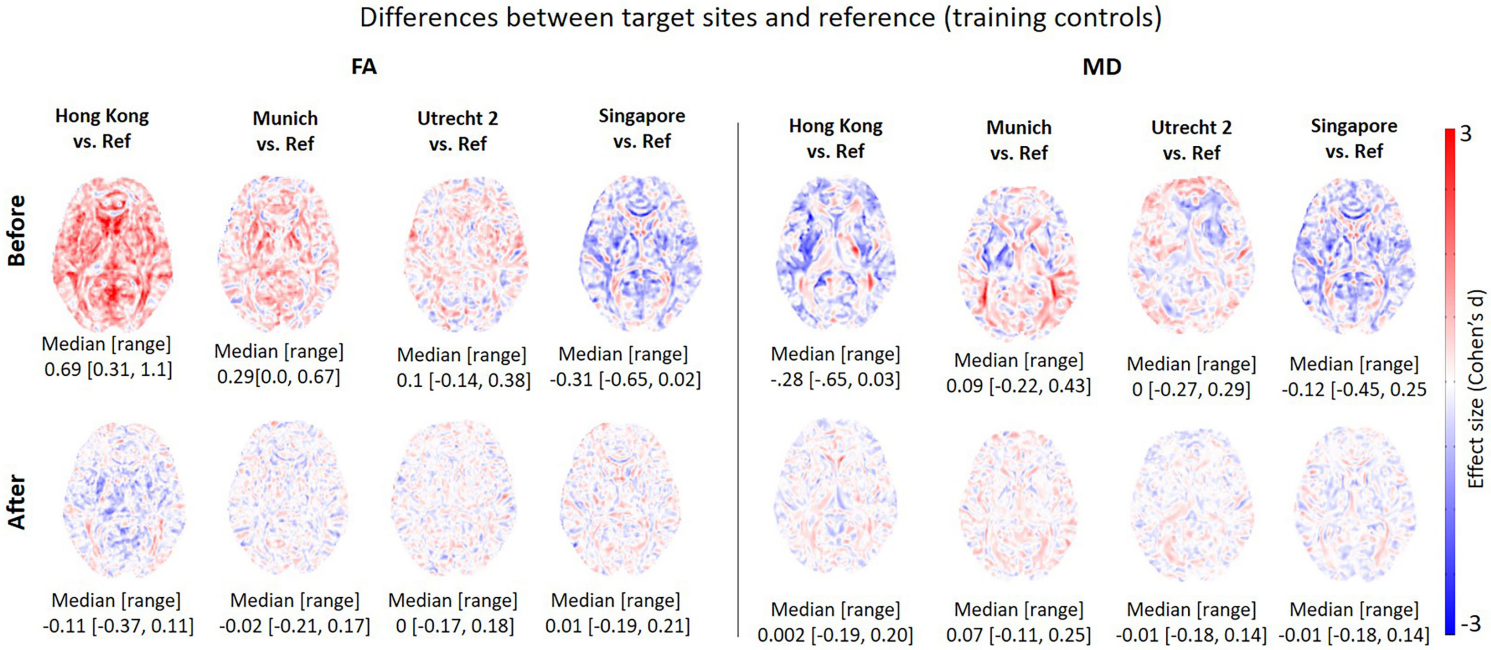


Figure S1: Results of whole brain voxel-wise analysis (right) comparing FA (top) and MD (bottom) between Training Controls of each target site and the reference, before and after harmonization. The red-blue colormap shows Cohen's d effect sizes. These maps support Figure 3, showing that effect sizes become close to zero after harmonization.

Part 4 - Sensitivity to disease effects: voxel-wise differences between patients and controls

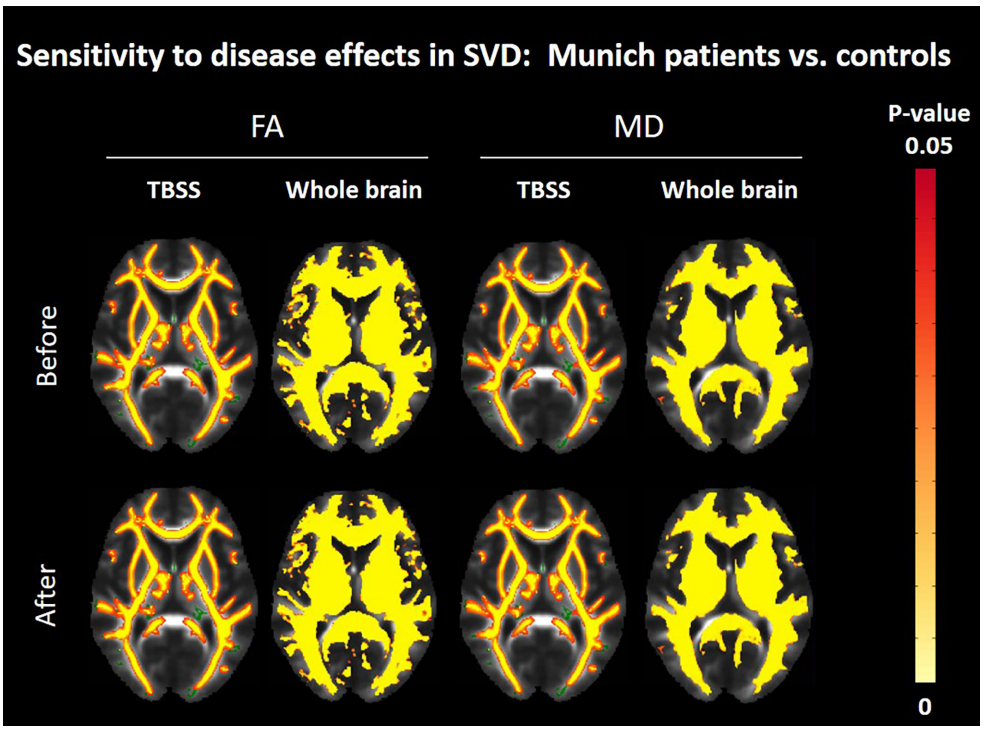


Figure S2: Voxel-wise differences in FA and MD between patients and controls (in Munich), before and after harmonization. Colormap shows $p < 0.05$

Part 5 - Effectiveness of harmonization using a second sample Training Controls

In Figure S1 we demonstrate that RISH harmonization is independent of the sample of training subjects used to train the algorithm, as long as these subjects are matched between the reference and the target. Differences between matched controls of Utrecht1 and Munich sites found before harmonization were removed after harmonization (note that the gray boxplots representing the controls fall on the same line after harmonization). Furthermore, within each site, effect sizes of FA, MD, PSMD found before harmonization were unaffected after harmonization. This result indicates that the RISH method is independent of sampling biases of Training Controls.

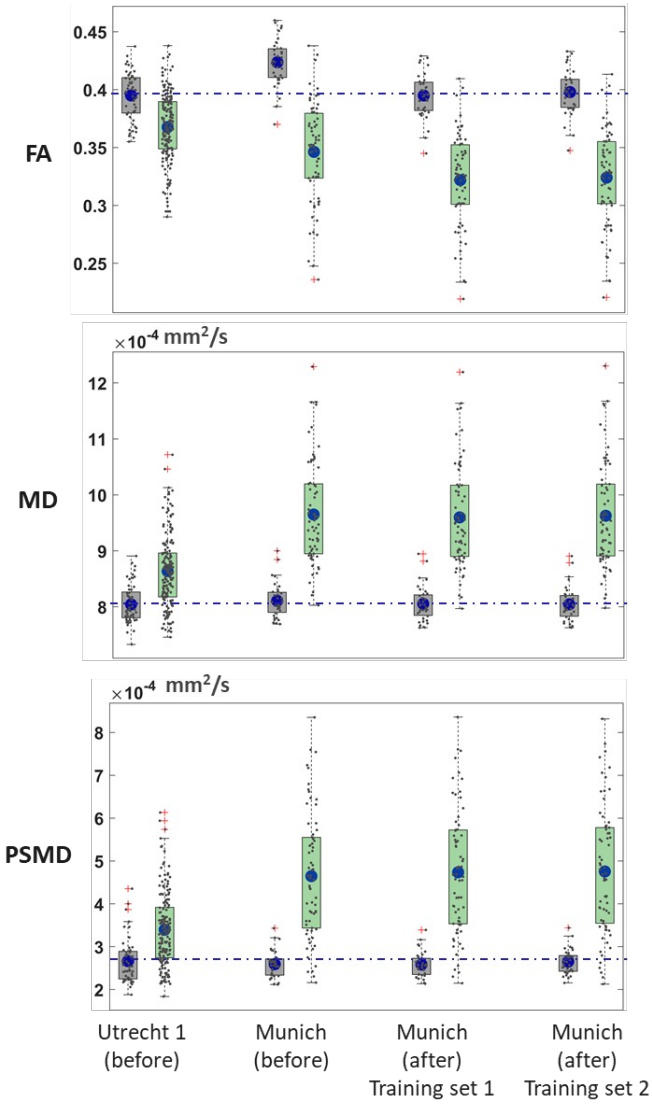


Figure S3: Average FA (top), MD (bottom) and PSMD of the white matter skeleton compared between patients (green) and controls (gray) within each site. Results are displayed for the reference site for the Munich site before and after harmonization using different training samples. Corresponding p-values and effect sizes are displayed in Table S4 The dashed line indicates the mean value of controls of the reference site and blue maker in each boxplot represent the mean of that group

Table S4: Effect sizes between patients and controls for the Munich site, before and after harmonizing the data with different sets of training controls.

		Utrecht1	Munich before	Munich After (Training 1)	% Change	Munich after (Training 2)	% Change
FA	P-value	1.4×10^{-11}	3.17×10^{-21}	2.30×10^{-21}	1	6×10^{-21}	0
	Effect size (Cohen's d)	-0.96	-2.07	-2.09		-2.07	
	CI	(-1.29, -.64)	(-2.57, -1.57)	(-2.59, -1.59)		(-2.51, -1.52)	
MD	P-value	6.29×10^{-10}	9.57×10^{-21}	6.29×10^{-21}	1	7.3×10^{-21}	1
	Effect size (Cohen's d)	1.02	1.99	2		2	
	CI	(.63, 1.34)	(1.49, 2.49)	(1.49, 2.48)		(1.49, 2.48)	
PSMD	P-value	3.146×10^{-12}	3.601×10^{-17}	5.971×10^{-18}	4	5.6×10^{-18}	4
	Effect size (Cohen's d)	0.93	1.71	1.78		1.77	
	CI	(.62, 1.25)	(1.23, 2.18)	(1.29, 2.25)		(1.29, 2.24)	

Part 6 - Statistical harmonization using ComBat

We compared the performance of RISH harmonization with a recently proposed statistical harmonization tool, ComBat (Fortin et al., 2017). We performed 3 experiments using different sets of covariates in the model. We used age and sex as biological covariates for all experiments. Our aim was to understand whether scanner related differences in matched controls between sites are removed, while preserving effect sizes between patients and controls within each site. In the first experiment, patients and controls were harmonized in the same batch using an additional covariate for "group" (e.g., patients = 2 vs. controls = 1). In the second experiment, controls and patients were also harmonized in the same batch, using a group covariate with values between 1 and 3 for controls, SVD patients and CADASIL patients, respectively. The aim of this analysis was to understand whether separating the subjects per disease type would enhance the performance of harmonization. Finally, in the third experiment we harmonized patients and controls in separate batches.

6.1 Experiment 1: using a group covariate of patient and controls.

Figure S2 and Table S5 show results of ComBat harmonization using disease type as covariate. After harmonization, one-way ANOVA shows that significant differences still exist between matched controls of all sites: FA ($F(3,120) = 31, p < 0.001$), MD ($F(3,120) = 40, p < 0.001$), and PSMD ($F(3,120) = 44, p < 0.001$). Note that the controls (gray boxplots) do not fall on the same line before or after harmonization, suggesting that scanner related differences are not accurately removed. Furthermore, effect sizes between patients and controls are not always preserved after harmonization, and changes in effect size up to 96% were observed.

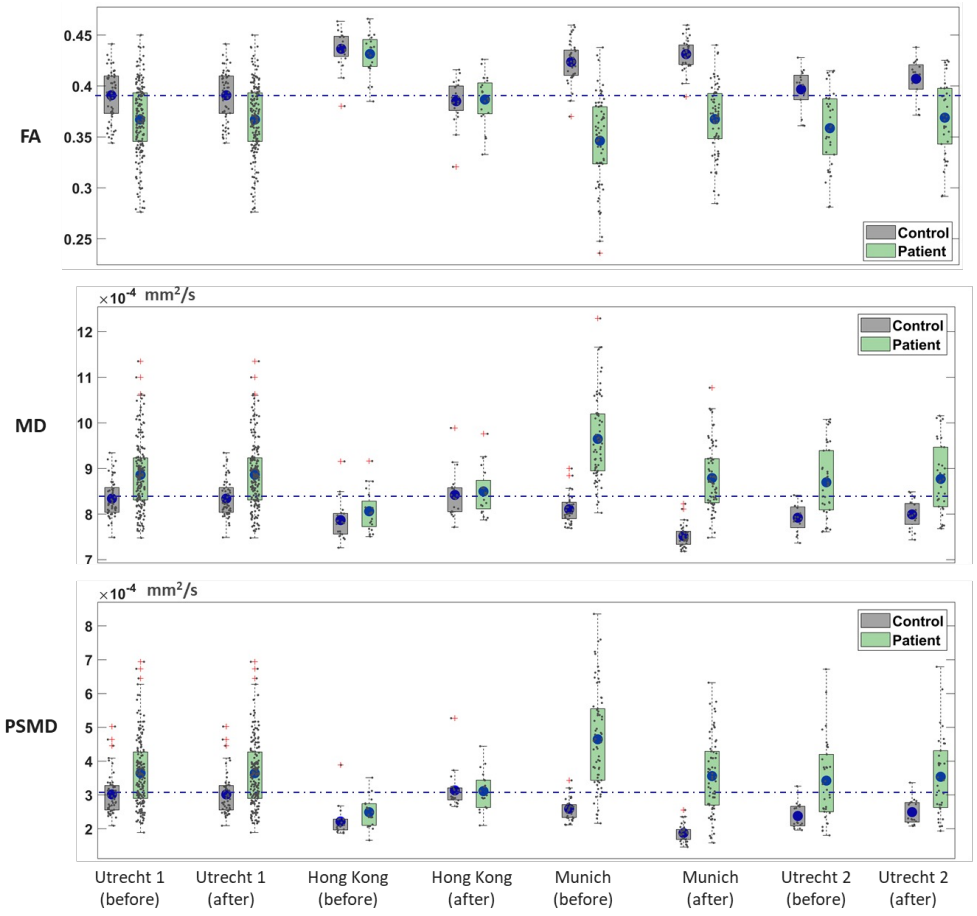


Figure S4: Harmonization using the ComBat method with age, sex and group (patients, controls) as covariates. Average FA (top), MD (bottom) and PSMD of the white matter skeleton compared between patients (green) and controls (gray) within each site, before and after harmonization. P-values and effect sizes are displayed in Table S4. The dashed line indicates the harmonized mean value of controls of the reference site and blue makers in each boxplot represents the mean of each group.

Table S5: Effect sizes between patients and controls within each site, before and after ComBat harmonization with age, sex and group (patients, controls) as covariates.

		Utrecht1 before	Utrecht1 after	% Change	Hong Kong before	Hong Kong after	% Change	Munich before	Munich after	% Change	Utrecht2 before	Utrecht2 after	% Change
FA	P-value	3.66×10^{-9}	9.00×10^{-6}		.43	.83		1.77×10^{-16}	5.36×10^{-18}		3.9×10^{-4}	3.7×10^{-4}	
	Effect size (Cohen's d)	-.97	-.71	27	-0.25	.07	72	-2.07	-2.21	-7	-1.16	-1.16	0
	CI (Cohen's d)	(-1.29, -.64)	(-1.02, -.40)		(-.87, .37)	(-.55, .69)		(-2.57, -1.57)	(-2.73, -1.70)		(-1.79, -.52)	(-1.79, -.52)	
MD	P-value	6.30×10^{-10}	2.00×10^{-6}		.17	.63		1.14×10^{-15}	5.16×10^{-17}		2.80×10^{-4}	2.9×10^{-4}	
	Effect size (Cohen's d)	1.01	.76	-25	.44	.15	-66	1.99	2.12	7	1.19	1.18	-1
	CI (Cohen's d)	(.69, 1.34)	(.45, 1.08)		(-.19, 1.06)	(-.47, .77)		(1.50, 2.49)	(1.61, 2.62)		(.54, 1.82)	(.54, 1.82)	
PSMD	P-value	1.12×10^{-8}	2.60×10^{-5}		.06	.91		1.09×10^{-12}	7.49×10^{-14}		.001	0.001	
	Effect size (Cohen's d)	.93	.68	-27	0.62	-.04	-94	1.71	1.82	6	1.03	1.65	60
	CI (Cohen's d)	(.61, 1.25)	(.36, .99)		(-.02, .25)	(-.66, .58)		(1.23, 2.18)	(1.33, 2.30)		(.39, 1.65)	(.41, 1.67)	

6.2 Experiment 2: using a group covariate of controls, SVD and CADASIL

Figure S3 and Table S6 show results of ComBat harmonization using disease type as covariate. Similar to the previous experiment, after harmonization, significant differences still persisted between matched controls of all sites: FA ($F(3,120) = 7.8$, $p < 0.001$), MD ($F(3,120) = 9.4$, $p < 0.001$), and PSMD ($F(3,120) = 5.9$, $p < 0.001$). Effect sizes between patients and controls were less affected by harmonization than in Experiment 1, showing that the inclusion of disease type as covariate can enhance the performance.

6.3 Experiment 3: patients and controls harmonized in separate batches.

Figure S4 and Table S7 shows results of ComBat harmonization when patients and controls are harmonized separately. Contrary to the previous experiments, after harmonization, differences in FA, MD and PSMD between matched controls of all sites were removed (all $p>0.05$). Note that control groups (gray boxes) fall on the same line after harmonization. However, effect sizes between patients and controls were severely disrupted by harmonization. For example, for the Hong site where FA differences were not initially significant (Cohen's = -0.25, p = .43), the effect size was inflated after harmonization (Cohen's = -2.11, $p = 6.74 \times 10\text{-}8$).

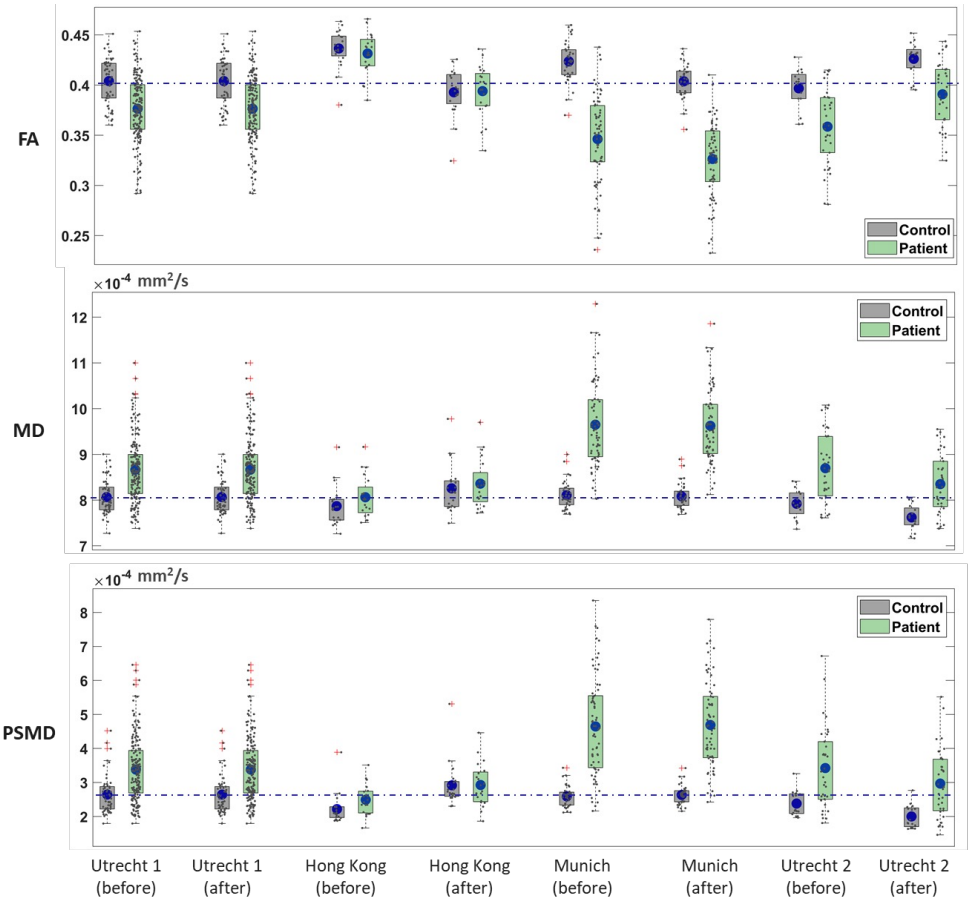


Figure S5: Harmonization using ComBat method with, age, sex and group (CADASIL, SVD, controls) as covariates. Average FA (top), MD (bottom) and PSMD of the white matter skeleton compared between patients (green) and controls (gray) within each site, before and after harmonization. P-values and effect sizes are displayed in Table S5. The dashed line indicates the harmonized mean value of controls of the reference site and blue makers in each boxplot represents the mean of each group.

Table S6: Effect sizes between patients and controls within each site, before and after ComBat harmonization with age, sex and group (CADASIL, SVD, controls) as covariates.

		Utrecht1 before	Utrecht1 after	% Change	Hong Kong before	Hong Kong after	% Change	Munich before	Munich after	% Change	Utrecht2 before	Utrecht2 after	% Change
FA	P-value	3.66×10^{-9}	4.88×10^{-8}		0.43	0.88		1.77×10^{-16}	1.70×10^{-19}		3.9×10^{-4}	1.40×10^{-4}	
	Effect size (Cohen's d)	-0.97	-0.89	8	-0.25	0.05	80	-2.07	-2.36	-14	-1.16	-1.26	-9
	CI (Cohen's d)	(-1.29, -.64)	(-1.21, -.57)		(-.87, .37)	(-.57, .67)		(-2.57, -1.57)	(-2.88, -1.83)		(-1.79, -.52)	(-1.89, -.61)	
MD	P-value	6.30×10^{-10}	1.01×10^{-8}		0.17	0.54		1.14×10^{-15}	1.38×10^{-18}		2.80×10^{-4}	1.02×10^{-4}	
	Effect size (Cohen's d)	1.01	0.94	-7	0.44	0.19	-57	1.99	2.27	14	1.19	1.28	8
	CI (Cohen's d)	(.69, 1.34)	(.62, 1.26)		(-.19, 1.06)	(-.43, .82)		(1.50, 2.49)	(1.75, 2.79)		(.54, 1.82)	(.63, 1.93)	
PSMD	P-value	1.12×10^{-8}	1.12×10^{-8}		0.06	0.97		1.09×10^{-12}	2.17×10^{-15}		0.001	5.41×10^{-4}	
	Effect size (Cohen's d)	0.93	0.85	-9	0.62	0.01	-98	1.71	1.97	15	1.03	1.12	9
	CI (Cohen's d)	(.61, 1.25)	(.36, .99)		(-.02, .25)	(-.66, .58)		(1.23, 2.18)	(1.33, 2.30)		(.39, 1.65)	(.41, 1.67)	

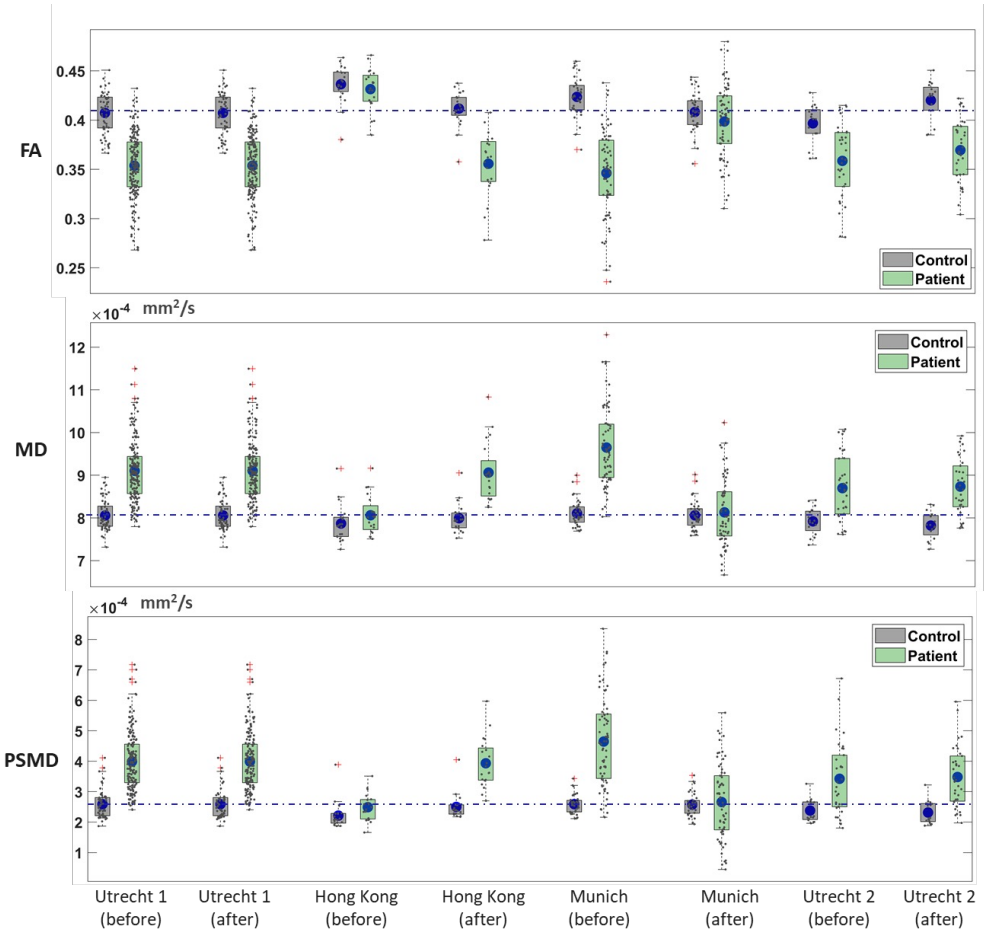


Figure S6: Harmonization using ComBat method, with patients and controls harmonized separately using age, sex and group as covariates. Average FA (top), MD (bottom) and PSMD of the white matter skeleton compared between patients (green) and controls (gray) within each site, before and after harmonization. P-values and effect sizes are displayed in Table S6. The dashed line indicates the harmonized mean value of controls of the reference site and blue makers in each boxplot represent the mean of each group.

Table S7: Effect sizes between patients and controls within each site, before and after ComBat harmonization with patients and controls harmonized separately. Age and sex were used as covariates.

		Utrecht1 before	Utrecht1 after	% Change	Hong Kong before	Hong Kong after	% Change	Munich before	Munich after	% Change	Utrecht2 before	Utrecht2 after	% Change
FA	P-value	3.66×10^{-9}	1.32×10^{-22}		0.43	6.74×10^{-8}		1.77×10^{-16}	0.15		3.9×10^{-4}	3.85×10^{-7}	
	Effect size (Cohen's d)	-0.97	-1.72	-77	-0.25	-2.11	-744	-2.07	-0.3	86	-1.16	-1.78	-53
	CI (Cohen's d)	(-1.29, -.64)	(-2.07, -1.37)		(-.87, .37)	(-2.88, -1.32)		(-2.57, -1.57)	(-.71, .11)		(-1.79, -.52)	(-2.47, -1.08)	
MD	P-value	6.30×10^{-10}	1.21×10^{-20}		0.17	3.17×10^{-7}		1.14×10^{-15}	0.61		2.80×10^{-4}	3.00×10^{-6}	
	Effect size (Cohen's d)	1.01	1.62	60	0.44	1.96	345	1.99	0.11	-94	1.19	1.61	35
	CI (Cohen's d)	(.69, 1.34)	(1.28, 1.96)		(-.19, 1.06)	(1.18, 2.71)		(1.50, 2.49)	(-.31, .52)		(.54, 1.82)	(.93, 2.28)	
PSMD	P-value	1.12×10^{-8}	2.19×10^{-20}		0.06	2.51×10^{-8}		1.09×10^{-12}	0.69		0.001	4.50×10^{-5}	
	Effect size (Cohen's d)	0.93	1.61	73	0.62	2.21	256	1.71	0.08	-95	1.03	1.36	32
	CI (Cohen's d)	(.61, 1.25)	(.36, .99)		(-.02, .25)	(-.66, .58)		(1.23, 2.18)	(1.33, 2.30)		(.39, 1.65)	(.41, 1.67)	

Part 7 - Prevalence of WMH in training controls.

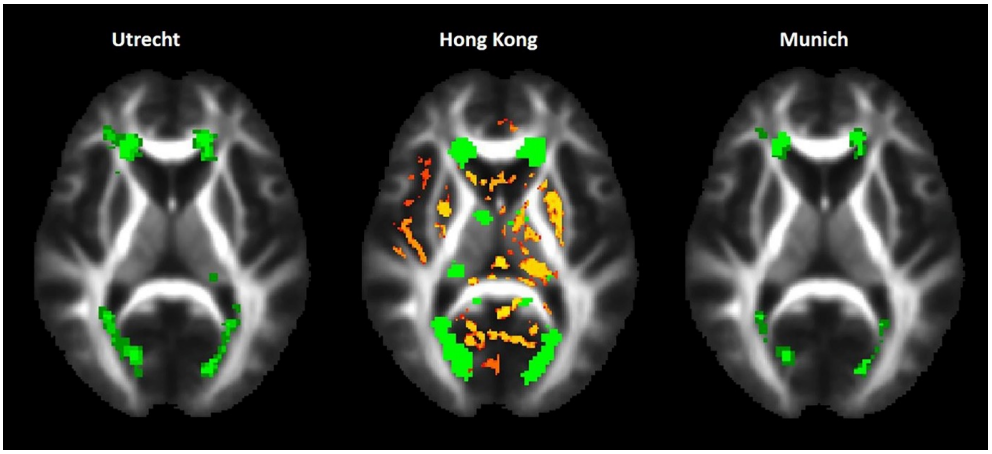


Figure S7: White matter hyperintensities (WMHs) observed in training controls of different sites. The green areas correspond to WMHs that occur in at least 25% of the subjects. For the Hong Kong site, the red-yellow colormap indicates areas where MD differences persisted after harmonization ($p<0.05$ when compared to the reference site).

REFERENCES

Aalten, P., Ramakers, I. H. G. B., Biessels, G. J., de Deyn, P. P., Koek, H. L., OldeRikkert, M. G. M., Oleksik, A. M., Richard, E., Smits, L. L., van Swieten, J. C., Teune, L. K., van der Lugt, A., Barkhof, F., Teunissen, C. E., Rozendaal, N., Verhey, F. R. J., & van der Flier, W. M. (2014). The Dutch Parelsnoer Institute - Neurodegenerative diseases; methods, design and baseline results. BMC Neurology, 14(1), 1–8. https://doi.org/10.1186/s12883-014-0254-4

Baykara, E., Gesierich, B., Adam, R., Tuladhar, A. M., Biesbroek, J. M., Koek, H. L., Ropele, S., Jouvent, E., Chabriat, H., Ertl-Wagner, B., Ewers, M., Schmidt, R., de Leeuw, F. E., Biessels, G. J., Dichgans, M., & Duering, M. (2016). A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. Annals of Neurology, 80(4), 581–592. https://doi.org/10.1002/ana.24758

Bendlin, B. B., Fitzgerald, M. E., Ries, M. L., Xu, G., Kastman, E. K., Thiel, B. W., Rowley, H. A., Lazar, M., Alexander, A. L., & Johnson, S. C. (2010). White matter in aging and cognition: A cross-sectional study of microstructure in adults aged eighteen to eighty-three. Developmental Neuropsychology, 35(3), 257–277. https://doi.org/10.1080/87565641003696775

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brink, H. Van Den, Kopczak, A., Arts, T., Onkenhout, L., Siero, J. C. W., Zwanenburg, J. J. M., Duering, M., Blair, G. W., Doubal, F. N., Stringer, M. S., Thrippleton, M. J., Kuijf, H. J., Luca, A. De, Hendrikse, J., Wardlaw, J. M., Dichgans, M., & Jan, G. (2021). Cerebral Circulation - Cognition and Behavior Zooming in on cerebral small vessel function in small vessel diseases with 7T MRI : Rationale and design of the " ZOOM @ SVDs " study. Cerebral Circulation - Cognition and Behavior, 2(April), 100013. https://doi.org/10.1016/j.cccb.2021.100013

Büchel, C., Raedler, T., Sommer, M., Sach, M., Weiller, C., & Koch, M. A. (2004). White matter asymmetry in the human brain: A diffusion tensor MRI study. Cerebral Cortex, 14(9), 945–951. https://doi.org/10.1093/cercor/bhh055

Cetin-Karayumak, S., Di Biase, M. A., Chunga, N., Reid, B., Somes, N., Lyall, A. E., Kelly, S., Solgun, B., Pasternak, O., Vangel, M., Pearlson, G., Tamminga, C., Sweeney, J. A., Clementz, B., Schretlen, D., Viher, P. V., Stegmayer, K., Walther, S., Lee, J., … Kubicki, M. (2020). White matter abnormalities across the lifespan of schizophrenia: a harmonized multi-site diffusion MRI study. Molecular Psychiatry, 25(12), 3208–3219. https://doi.org/10.1038/s41380-019-0509-y

Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., & Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. NeuroImage, 184, 180–200. https://doi.org/10.1016/j.neuroimage.2018.08.073

De Luca, A., Guo, F., Froeling, M., & Leemans, A. (2020). Spherical deconvolution with tissue-specific response functions and multi-shell diffusion MRI to estimate multiple fiber orientation distributions (mFODs). NeuroImage, 222(January), 117206. https://doi.org/10.1016/j.neuroimage.2020.117206

De Luca, A., Leemans, A., Bertoldo, A., Arrigoni, F., & Froeling, M. (2018). A robust deconvolution method to disentangle multiple water pools in diffusion MRI. NMR in Biomedicine, 31(11), 1–17. https://doi.org/10.1002/nbm.3965

Du, J., Zhu, H., Zhou, J., Lu, P., Qiu, Y., Yu, L., Cao, W., Zhi, N., Yang, J., Xu, Q., Sun, J., & Zhou, Y. (2020). Structural Brain Network Disruption at Preclinical Stage of Cognitive Impairment Due to Cerebral Small Vessel Disease. Neuroscience, 449, 99–115. https://doi.org/10.1016/j.neuroscience.2020.08.037

Fazekas, F., Chawluk, J. B., & Alavi, A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. American Journal of Neuroradiology, 8(3), 421–426.

Finsterwalder, S., Vlegels, N., Gesierich, B., Araque Caballero, M., Weaver, N. A., Franzmeier, N., Georgakis, M. K., Konieczny, M. J., Koek, H. L., Karch, C. M., Graff-Radford, N. R., Salloway, S., Oh, H., Allegri, R. F., Chhatwal, J. P., Jessen, F., Düzel, E., Dobisch, L., Metzger, C., … Duering, M. (2020). Small vessel disease more than Alzheimer's disease determines diffusion MRI alterations in memory clinic patients. Alzheimer's and Dementia, March, 1504–1514. https://doi.org/10.1002/alz.12150

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. NeuroImage, 161(July), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Helmer, K. G., Chou, M.-C., Preciado, R. I., Gimi, B., Rollins, N. K., Song, A., Turner, J., & Mori, S. (2016). Multi-site study of diffusion metric variability: effects of site, vendor, field strength, and echo time on regions-of-interest and histogram-bin analyses. Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging, 9788, 97882U. https://doi.org/10.1117/12.2217445

Hilal, S., Ikram, M. K., Saini, M., Tan, C. S., Catindig, J. A., Dong, Y. H., Lim, L. B. S., Ting, E. Y. S., Koo, E. H., Cheung, C. Y. L., Qiu, A., Wong, T. Y., Chen, C. L. H., & Venketasubramanian, N. (2013). Prevalence of cognitive impairment in Chinese: Epidemiology of Dementia in Singapore study. Journal of Neurology, Neurosurgery and Psychiatry, 84(6), 686–692. https://doi.org/10.1136/jnnp-2012-304080

Hilal, S., Liu, S., Wong, T. Y., Vrooman, H., Cheng, C. Y., Venketasubramanian, N., Chen, C. L. H., & Zhou, J. H. (2021). White matter network damage mediates association between cerebrovascular disease and cognition. Journal of Cerebral Blood Flow and Metabolism. https://doi.org/10.1177/0271678X21990980

Hsu, J.L., Leemans, A., Bai, C.H., Lee, C.H., Tsai, Y.F., Chiu, H.C., Chen, W.H., (2008). Gender differences and age-related white matter changes of the human brain: A diffusion tensor imaging study. Neuroimage 39, 566–577. doi: 10.1016/j. neuroimage.2007.09.017 .

Konieczny, M. J., Dewenter, A., Telgte, A. ter, Gesierich, B., Wiegertjes, K., Finsterwalder, S., Kopczak, A., Hübner, M., Malik, R., Tuladhar, A. M., Marques, J. P., Norris, D. G., Koch, A., Dietrich, O., Ewers, M., Schmidt, R., de Leeuw, F.-E., & Duering, M. (2020). Multi-shell diffusion MRI models for white matter characterization in cerebral small vessel disease. In Neurology. https://doi.org/10.1212/wnl.0000000000011213

Koppers, S., Bloy, L., Berman, J. I., Tax, C. M. W., Edgar, J. C., & Merhof, D. (2019). Spherical Harmonic Residual Network for Diffusion Signal Harmonization. Mathematics and Visualization, 226249, 173–182. https://doi.org/10.1007/978-3-030-05831-9_14

Kumar, R., Husain, M., Gupta, R. K., Hasan, K. M., Haris, M., Agarwal, A. K., Pandey, C. M., & Narayana, P. A. (2009). Serial changes in the white matter diffusion tensor imaging metrics in moderate traumatic brain injury and correlation with neuro-cognitive function. Journal of Neurotrauma, 26(4), 481–495. https://doi.org/10.1089/neu.2008.0461

Lam, B. Y. K., Leung, K. T., Yiu, B., Zhao, L., Biesbroek, J. M., Au, L., Tang, Y., Wang, K., Fan, Y., Fu, J. H., Xu, Q., Song, H., Tian, X., Chu, W. C. W., Abrigo, J., Shi, L., Ko, H., Lau, A., Duering, M., … Mok, V. C. T. (2019). Peak width of skeletonized mean diffusivity and its association with age-related cognitive alterations and vascular risk factors. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, 11, 721–729. https://doi.org/10.1016/j.dadm.2019.09.003

Leijsen, E. M. C. Van, Uden, I. W. M. Van, Bergkamp, M. I., Holst, H. M. Van Der, Norris, D. G., Claassen, J. A. H. R., & Kessels, R. P. C. (2019). NeuroImage : Clinical Longitudinal changes in rich club organization and cognition in cerebral small vessel disease. NeuroImage: Clinical, 24(October), 102048. https://doi.org/10.1016/j.nicl.2019.102048

Lorca-Puls, D. L., Gajardo-Vidal, A., White, J., Seghier, M. L., Leff, A. P., Green, D. W., Crinion, J. T., Ludersdorfer, P., Hope, T. M. H., Bowman, H., & Price, C. J. (2018). The impact of sample size on the reproducibility of voxel-based lesion-deficit mappings. Neuropsychologia, 115(June 2017), 101–111. https://doi.org/10.1016/j.neuropsychologia.2018.03.014

Lyoubi-Idrissi, A. L., Jouvent, E., Poupon, C., & Chabriat, H. (2017). Diffusion magnetic resonance imaging in cerebral small vessel disease. Revue Neurologique, 173(4), 201–210. https://doi.org/10.1016/j.neurol.2017.03.005

Magin RL, Karaman MM, Hall MG, Zhu W, Zhou XJ. (2019). Capturing complexity of the diffusion-weighted MR signal decay. Magn Reson Imaging. Feb;56:110-118. doi: 10.1016/j.mri.2018.09.034. Epub 2018 Oct 9. PMID: 30314665; PMCID: PMC6348133.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., … Mazoyer, B. (2001). A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philosophical Transactions of the Royal Society B: Biological Sciences, 356(1412), 1293–1322. https://doi.org/10.1098/rstb.2001.0915

Mirzaalian, H., De Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M., Westin, C. F., Shenton, M. E., & Rathi, Y. (2015). Harmonizing diffusion MRI data across multiple sites and scanners. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9349, 12–19. https://doi.org/10.1007/978-3-319-24553-9_2

Ning, L., Bonet-Carne, E., Grussu, F., Sepehrband, F., Kaden, E., Veraart, J., Blumberg, S. B., Khoo, C. S., Palombo, M., Kokkinos, I., Alexander, D. C., Coll-Font, J., Scherrer, B., Warfield, S. K., Karayumak, S. C., Rathi, Y., Koppers, S., Weninger, L., Ebert, J., … Tax, C. M. W. (2020). Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: Algorithms and results. NeuroImage, 221(October 2019). https://doi.org/10.1016/j.neuroimage.2020.117128

Palacios, E. M., Martin, A. J., Boss, M. A., Ezekiel, F., Chang, Y. S., Yuh, E. L., Vassar, M. J., Schnyer, D. M., MacDonald, C. L., Crawford, K. L., Irimia, A., Toga, A. W., & Mukherjee, P. (2017). Toward precision and reproducibility of diffusion tensor imaging: A multicenter diffusion phantom and traveling volunteer study. American Journal of Neuroradiology, 38(3), 537–545. https://doi.org/10.3174/ajnr.A5025

Reijmer, Y. D., Fotiadis, P., Martinez-Ramirez, S., Salat, D. H., Schultz, A., Shoamanesh, A., Ayres, A. M., Vashkevich, A., Rosas, D., Schwab, K., Leemans, A., Biessels, G. J., Rosand, J., Johnson, K. A., Viswanathan, A., Gurol, M. E., & Greenberg, S. M. (2015). Structural network alterations and neurological dysfunction in cerebral amyloid angiopathy. Brain, 138(1), 179–188. https://doi.org/10.1093/brain/awu316

Reijmer, Y. D., Leemans, A., Brundel, M., Kappelle, L. J., & Biessels, G. J. (2013). Disruption of the cerebral white matter network is related to slowing of information processing speed in patients with type 2 diabetes. Diabetes, 62(6), 2112–2115. https://doi.org/10.2337/db12-1644

Rydhög, A. S., Szczepankiewicz, F., Wirestam, R., Ahlgren, A., Westin, C. F., Knutsson, L., & Pasternak, O. (2017). Separating blood and water: Perfusion and free water elimination from diffusion MRI in the human brain. NeuroImage, 156(December 2016), 423–434. https://doi.org/10.1016/j.neuroimage.2017.04.023

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., & Behrens, T. E. J. (2006). Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. NeuroImage, 31(4), 1487–1505. https://doi.org/10.1016/j.neuroimage.2006.02.024

Steven AJ, Zhuo J, Melhem ER, (2014). Diffusion kurtosis imaging: an emerging technique for evaluating the microstructural environment of the brain. AJR. American journal of roentgenology 202 (1), W26–W33

Takao, H., Hayashi, N., Kabasawa, H., & Ohtomo, K. (2012). Effect of scanner in longitudinal diffusion tensor imaging studies. Human Brain Mapping, 33(2), 466–477. https://doi.org/10.1002/hbm.21225

Tax, C. M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., John Evans, C., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., Ghosh, A., Tanno, R., Alexander, D. C., Zappalà, S., Charron, C., Kusmia, S., Linden, D. E., Jones, D. K., & Veraart, J. (2019). Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. NeuroImage, 195(August 2018), 285–299. https://doi.org/10.1016/j.neuroimage.2019.01.077

Teipel, S. J., Reuter, S., Stieltjes, B., Acosta-Cabronero, J., Ernemann, U., Fellgiebel, A., Filippi, M., Frisoni, G., Hentschel, F., Jessen, F., Klöppel, S., Meindl, T., Pouwels, P. J. W., Hauenstein, K. H., & Hampel, H. (2011). Multicenter stability of diffusion tensor imaging measures: A European clinical and physical phantom study. Psychiatry Research - Neuroimaging, 194(3), 363–371. https://doi.org/10.1016/j.pscychresns.2011.05.012

Timmermans, C., Smeets, D., Verheyden, J., Terzopoulos, V., Anania, V., Parizel, P. M., & Maas, A. (2019). Potential of a statistical approach for the standardization of multicenter diffusion tensor data: A phantom study. Journal of Magnetic Resonance Imaging, 49(4), 955–965. https://doi.org/10.1002/jmri.26333

Tournier, J. D., Calamante, F., Gadian, D. G., & Connelly, A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. NeuroImage, 23(3), 1176–1185. https://doi.org/10.1016/j.neuroimage.2004.07.037

Tuladhar, A. M., Van Norden, A. G. W., De Laat, K. F., Zwiers, M. P., Van Dijk, E. J., Norris, D. G., & De Leeuw, F. E. (2015). White matter integrity in small vessel disease is related to cognition. NeuroImage: Clinical, 7, 518–524. https://doi.org/10.1016/j.nicl.2015.02.003

Van Horn, J. D., & Toga, A. W. (2014). Human neuroimaging as a "Big Data" science. Brain Imaging and Behavior, 8(2), 323–331. https://doi.org/10.1007/s11682-013-9255-y

Van Leijsen, E. M. C., Van Uden, I. W. M., Ghafoorian, M., Bergkamp, M. I., Lohner, V., Kooijmans, E. C. M., Van Der Holst, H. M., Tuladhar, A. M., Norris, D. G., Van DIjk, E. J., Rutten-Jacobs, L. C. A., Platel, B., Klijn, C. J. M., & De Leeuw, F. E. (2017). Nonlinear temporal dynamics of cerebral small vessel disease. Neurology, 89(15), 1569–1577. https://doi.org/10.1212/WNL.0000000000004490

Van Norden, A. G. W., De Laat, K. F., Van Dijk, E. J., Van Uden, I. W. M., Van Oudheusden, L. J. B., Gons, R. A. R., Norris, D. G., Zwiers, M. P., & De Leeuw, F. E. (2012). Diffusion tensor imaging and cognition in cerebral small vessel disease. The RUN DMC study. Biochimica et Biophysica Acta - Molecular Basis of Disease, 1822(3), 401–407. https://doi.org/10.1016/j.bbadis.2011.04.008

Veraart, J., Sijbers, J., Sunaert, S., Leemans, A., & Jeurissen, B. (2013). Weighted linear least squares estimation of diffusion MRI parameters: Strengths, limitations, and pitfalls. NeuroImage, 81, 335–346. https://doi.org/10.1016/j.neuroimage.2013.05.028

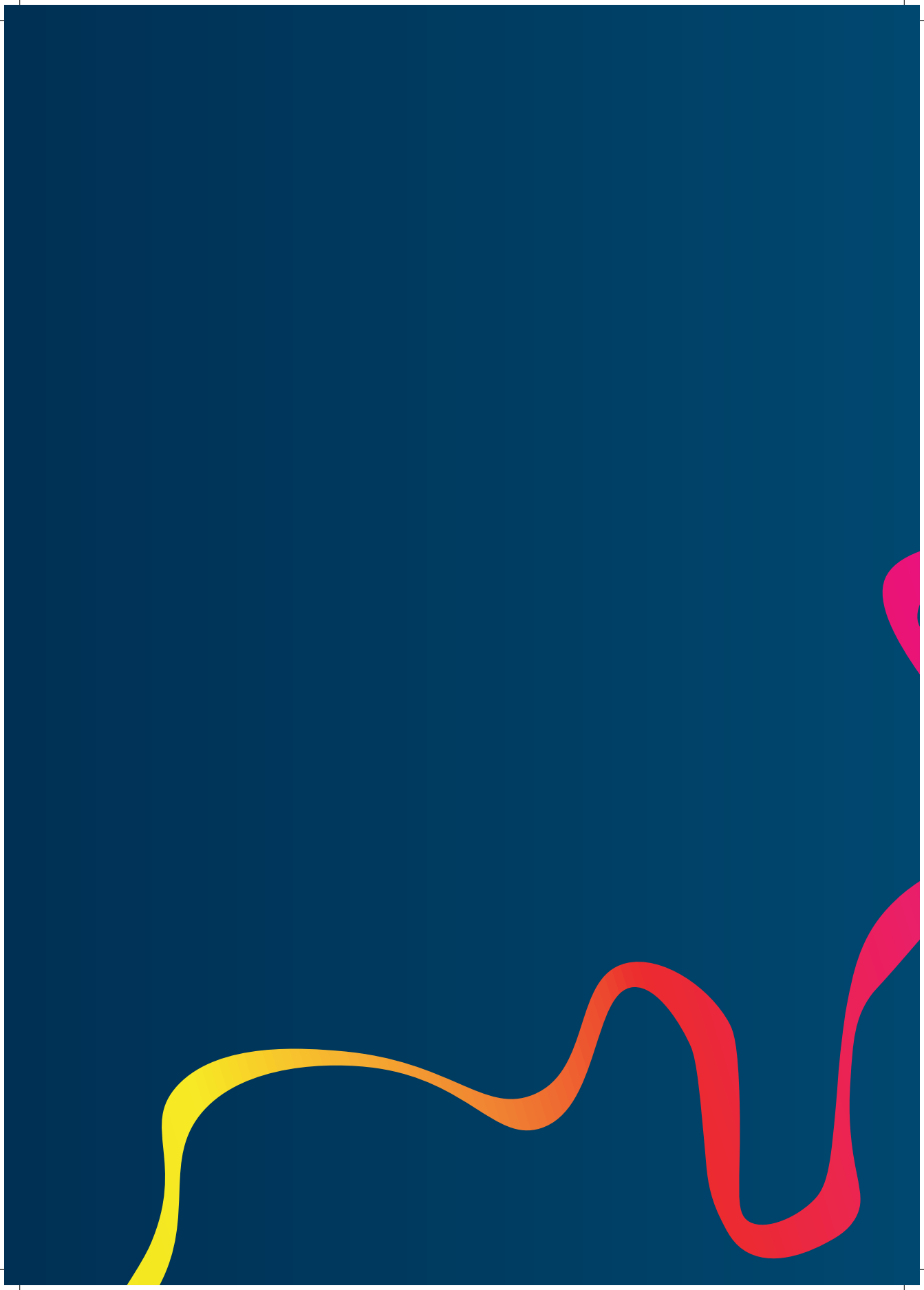
Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P., & Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. NeuroImage, 51(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Vos, S. B., Jones, D. K., Jeurissen, B., Viergever, M. A., & Leemans, A. (2012). The influence of complex white matter architecture on the mean diffusivity in diffusion tensor MRI of the human brain. NeuroImage, 59(3), 2208–2216. https://doi.org/10.1016/j.neuroimage.2011.09.086

Vos, S. B., Jones, D. K., Viergever, M. A., & Leemans, A. (2011). Partial volume effect as a hidden covariate in DTI analyses. NeuroImage, 55(4), 1566–1576. https://doi.org/10.1016/j.neuroimage.2011.01.048

Wiegertjes, K., Ter Telgte, A., Oliveira, P. B., Van Leijsen, E. M. C., Bergkamp, M. I., Van Uden, I. W. M., Ghafoorian, M., Van Der Holst, H. M., Norris, D. G., Platel, B., Klijn, C. J. M., Tuladhar, A. M., & De Leeuw, F. E. (2019). The role of small diffusion-weighted imaging lesions in cerebral small vessel disease. Neurology, 93(17), E1627–E1634. https://doi.org/10.1212/WNL.0000000000008364

Wollenweber, F. A., Hanecker, P., Bayer-Karpinska, A., Malik, R., Bäzner, H., Moreton, F., Muir, K. W., Müller, S., Giese, A., Opherk, C., Dichgans, M., Haffner, C., & Duering, M. (2015). Cysteine-sparing CADASIL mutations in NOTCH3 show proaggregatory properties in vitro. Stroke, 46(3), 786–792. https://doi.org/10.1161/STROKEAHA.114.007472



Chapter 3

Effect of fixed-density thresholding on structural brain networks: a demonstration in cerebral small vessel disease

Bruno M. de Brito Robalo
Naomi Vlegels
Jil Meier
Alexander Leemans
Geert Jan Biessels
Yael D. Reijmer

Brain Connectivity 2020 Apr, 10: 121-133
DOI: 10.1089/brain.2019.0686

ABSTRACT

Introduction: A popular solution to control for edge density variability in structural brain network analysis is to threshold the networks to a fixed density across all subjects. However, it remains unclear how this type of thresholding affects the basic network architecture in terms of edge weights, hub location, and hub connectivity and, especially, how it affects the sensitivity to detect disease-related abnormalities. We investigated these two questions in a cohort of patients with cerebral small vessel disease and age-matched controls.

Methods: Brain networks were reconstructed from diffusion magnetic resonance imaging data using deterministic fibre tractography. Networks were thresholded to a fixed density by removing edges with the lowest number of streamlines. We compared edge length (mm), fractional anisotropy (FA), proportion of hub connections, and hub location between the unthresholded and the thresholded networks of each subject. Moreover, we compared weighted graph measures of global and local connectivity obtained from the (un)thresholded networks between patients and controls. We performed these analyses over a range of densities (2–20%).

Results: Results indicate that fixed-density thresholding disproportionally removes edges composed of long streamlines but is independent of FA. The edges removed were not preferentially connected to hub or non-hub nodes. Over half of the original hubs were reproducible when networks were thresholded to a density ≥10%. Furthermore, the between-group differences in graph measures observed in the unthresholded network remained present after thresholding, irrespective of the chosen density.

Conclusions: We therefore conclude that moderate fixed-density thresholds can successfully be applied to control for the effects of density in structural brain network analysis.

1 INTRODUCTION

Studies of brain network changes in disease have improved our understanding of a wide range of cognitive and neurological disorders (Bassett and Bullmore, 2009; Stam, 2014). In network neuroscience, the structural brain network is often described as a set of nodes (gray matter regions) connected by edges (white matter tracts). Then, graph theoretical analysis is applied to characterize this network, allowing us to examine topological differences between patients and controls (Liu et al., 2017; Rubinov and Sporns, 2010; Sporns et al., 2005).

Although methods to reconstruct brain networks are improving, several methodological issues remain (Qi et al., 2015; Stam, 2014). A well-known issue is the confounding effect of network density (i.e., the number of detected edges divided by the number of all possible edges) on graph metrics (Stam, 2014; van Wijk et al., 2010). Since graph metrics are mathematically dependent on network density, variation in density caused by factors unrelated to disease of interest can mask real group differences or introduce false group differences (van Wijk et al., 2010). Such unwanted variation can be caused by factors related to network reconstruction. For example, brain networks reconstructed from diffusion magnetic resonance imaging (MRI) data contain false-positive and false-negative connections due to errors in fibre tractography, which occur especially in white matter regions with complex fibre orientations (Hein et al., 2016; Jeurissen et al., 2017). These complex fibre orientations cannot be fully resolved by using the diffusion tensor model, resulting in early termination of fibre tracts (i.e., false negatives) (Jones, 2008). On the contrary, advanced diffusion models such as constrained spherical deconvolution (CSD) are more effective in resolving complex fibre orientations but are prone to detect many false positives (Zalesky et al., 2016). Altogether, these reconstruction errors contribute to the variation in network density across subjects, affecting the computation and comparison of graph metrics.

One of the most straightforward methods to control for the variability in network density is to threshold the connectivity matrix by removing edges with the lowest weight until a common fixed density is achieved across all individuals (Achard and Bullmore, 2007; Bassett et al., 2009; van den Heuvel et al., 2008). Fixed-density thresholding is widely used in functional connectivity studies and has been suggested to be more reliable than other thresholding approaches such as absolute thresholding (Mårtensson et al., 2018; Telesford et al., 2011). However, the advantages and shortcomings of fixed-density thresholding are still under discussion (Garrison et al., 2015; van den Heuvel et al., 2017; Váša et al., 2018).

Much less is known about the consequences of fixed-density thresholding on diffusion-based brain networks (Andreotti et al., 2014b; Drakesmith et al., 2015). In particular, the effect of thresholding on the basic architecture of diffusion-based networks (e.g., edge composition, location of hub nodes and hub connections) is still unclear. The influence of thresholding on these network properties is relevant since these characteristics are the building blocks of structural networks and are often the focus of clinical studies (Achard et al., 2012; Daianu et al., 2016; Pandit et al., 2013). For example, long-range fibre connections are suggested to be more affected in certain

neurological disorders (Daianu et al., 2015). Moreover, network hubs and hub connections are preferentially disrupted in many pathologies and are frequently the most discriminatory features between patients and controls (Fagerholm et al., 2015; Tuladhar et al., 2017). Thus, before applying fixed-density thresholding in clinical studies, it is essential to understand if and how it affects the edge composition, the organization of hubs and hub connections in the network as well as the subsequent between-group comparisons of graph metrics.

In this study, we assessed the effects of thresholding on (1) edge composition, location of hub nodes and hub connectivity and (2) the ability to detect between-group differences in a cohort of patients with cerebral small vessel disease (SVD). SVD is a common cause of white matter injury in the elderly, and previous studies have characterized network abnormalities in patients with SVD based on unthresholded networks (Heinen et al., 2018; Reijmer et al., 2015). We evaluated our research objectives over a range of network densities while maintaining a fully connected network (van Dellen et al., 2018).

2 MATERIALS AND METHODS

2.1 Participants

Patients included in this study were part of the TRACE-VCI cohort recruited through the memory clinic at the University Medical Center Utrecht (Boomsma et al., 2017). Inclusion and exclusion criteria of the original cohort are described elsewhere (Boomsma et al., 2017; Heinen et al., 2018). Patients were eligible for this study if they had (1) a structural brain MRI scan, including a diffusion-weighted sequence; (2) evidence of SVD on MRI, such as white matter hyperintensities (WMH) and lacunes; and (3) if they presented with cognitive complaints, resulting in 162 patients.

In addition, 53 age-, sex-, and education-matched elderly without any history of dementia or other known neurological disorder that could affect cognition, and a minimental state examination (MMSE) score between 28 and 30, were included as a control group (Table 1). Control participants were recruited from the population as part of the Utrecht Diabetic Encephalopathy Study (van Den Berg et al., 2010).

2.2 Magnetic resonance imaging data acquisition

MRI data of patients and controls were acquired on a 3-T Philips scanner (Achieva, Philips, Best, the Netherlands) using a standardized protocol that included a 3D T1-weighted image and a diffusion-weighted sequence. The T1 scans were acquired with the following parameters: 192 continuous slices, isotropic acquisition voxel size: $1 \times 1 \times 1\ mm^3$, a flip angle of 8°, echo time (TE): 4.5 ms and repetition time (TR): 7.9 ms.

The diffusion-weighted scans were obtained using a single-shot spin echo EPI sequence with

Table 1: Characteristics of the Study Population

	SVD patients N = 137	Controls N = 44
Demographic characteristics		
Female sex	65 (47.4)	21 (47.7)
Age in years	74.01 ± 7.88	72.0 ± 4.6
MMSE	26 [23, 28]	29 [28, 30]*
Imaging markers		
Brain volume in % of ICV	67.5 [64.69, 69.3)	70.9.18 [69.9, 73.0]*
GM volume in % of ICV	35.4 [34.1, 36.7]	38.4 [36.5, 39.7]
WMH volume in % of ICV	1.1 [0.4, 2.2]	0.4 [0.2, 0.8]*
Lacunar infarct	46 (34)	13 (30)

Data presented as mean ± standard deviation, absolute number (%), or median [interquartile range]. Demographic variables and MRI markers of SVD were compared between patients and controls using independent-samples t-test for continuous variables and $\chi 2$ test for proportions. For the MMSE, brain volume, GM volume, and WMH volume, we used the nonparametric equivalent of the independent-samples t-test, Mann–Whitney U test, since these variables were not normally distributed. *$p < 0.02$. GM, gray matter; ICV, intracranial volume; MMSE, minimental state examination; MRI, magnetic resonance imaging; SVD, small vessel disease; WMH, white matter hyperintensities.

48 contiguous slices, field of view of $220 \times 220 \times 120\ mm^3$, isotropic acquisition voxel size: $2.5 \times 2.5 \times 2.5\ mm^3$, TR/TE 6638/73 ms, 45 isotopically distributed diffusion-sensitizing gradients with a b-value of 1200 s/mm2, and 1 b = 0 s/mm^2 (3 signal averages).

Furthermore, a fluid-attenuated inversion recovery sequence (FLAIR; TR/TE/Inversion time = 11,000/125/2800 ms) was acquired to rate neuroimaging markers of SVD (i.e., brain volumes, WMH volume, and lacunes).

2.3 Diffusion MRI processing and fibre tractography

The diffusion-weighted data were processed using ExploreDTI version 4.8.6 (Leemans et al., 2009) running on MATLAB R2014b (MATLAB and Statistics Toolbox Release 2014b; The MathWorks, Inc., Natick, MA). Data were corrected for signal drift (Vos et al., 2017), subject motion, eddy current, and susceptibility artifacts, including rotation of the B-matrix before the estimation of the diffusion tensors (Leemans and Jones, 2009; Tax et al., 2015; Veraart et al., 2013).

The diffusion tensors were computed using robust estimators (Tax et al., 2015) followed by whole-brain tractography. Fibre tracts were reconstructed by starting seed points uniformly throughout the data at 2 mm isotropic resolution with a step size of 1 mm. Each streamline was propagated using integration over fibre orientation distributions (FODs). Streamlines were guided by fibre orientations inferred using CSD with a maximum harmonic order (l-max) of 6. This method allows for the reconstruction of more complex pathways, such as those found in regions of crossing fibres (Jeurissen et al., 2011). Streamlines were terminated when they entered a voxel with FOD <0.1, or when the deflection angle between two successive steps was >45°. These tractography parameters were chosen according to recommendations from the previous literature (Jeurissen et al., 2017).

2.4 Network reconstruction

The T1-weighted images were pre-processed using the Computational Anatomy Toolbox (CAT12) from SPM. The pipeline, as illustrated in Figure 1, included brain extraction and parcellation of the brain volume into 90 cortical and subcortical regions of interest (ROIs), using the automated anatomical labelling (AAL) template (Tzourio-Mazoyer et al., 2002). The parcellation was performed in the native space, with the AAL template being warped to each subject's T1 image (Odish et al., 2015). After parcellation, we visually checked the quality of each image, and overlaid the segmented AAL labels with the original T1 image to ensure that the registration and parcellation were successful. The cerebellum was excluded from the analysis since most subjects had insufficient scan coverage of this structure.

For network definition, each ROI represented a node in the network. Two nodes were considered to be connected if they contained the endpoints of at least one streamline, resulting in a 90×90 binary connectivity matrix. We also obtained three weighted connectivity matrices for each subject, in which the previously defined edges were weighted by the number of streamlines

(NOS), average streamline length (SL), and fractional anisotropy (FA) (Figure 1).

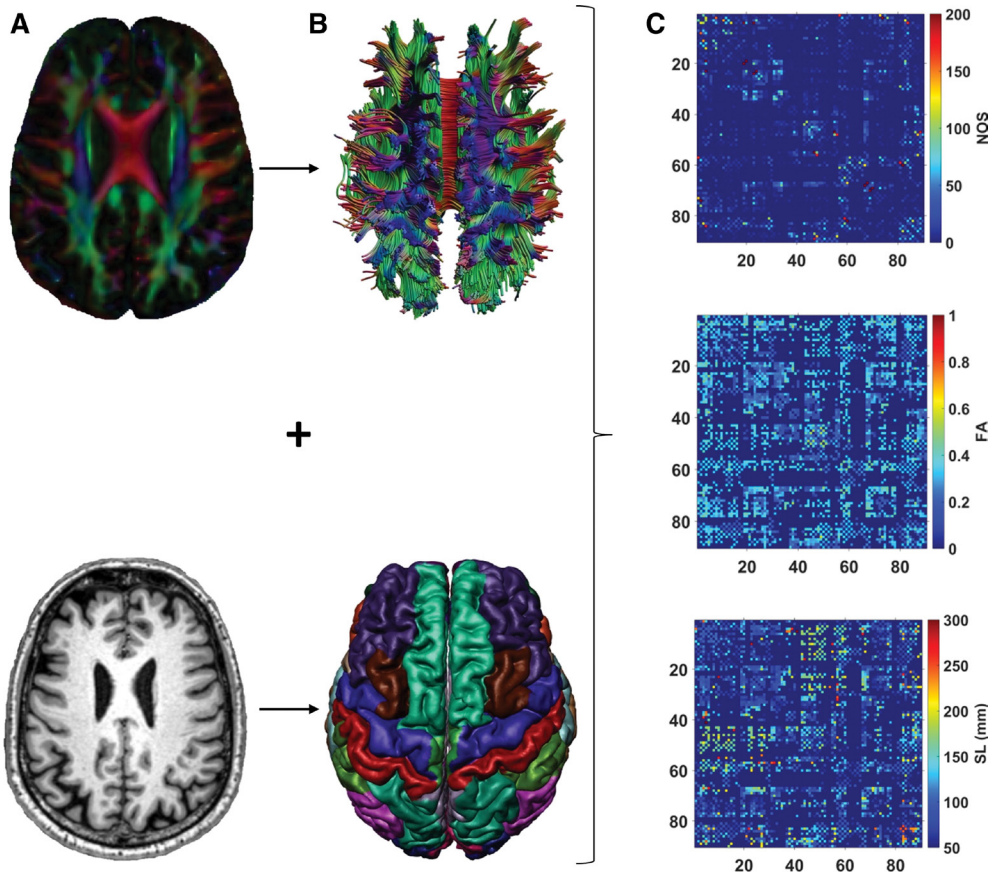


Figure 1: Flowchart of the structural network reconstruction. A: Coregistration of the diffusion-weighted image (top) with the T1-weighted image (bottom). B: Reconstruction of fibre tracts from the diffusion-weighted data (top) and the parcellation of the T1 image into 90 cortical and subcortical regions using AAL template (bottom). C: Network definition: for any pair of ROIs of the AAL template, we established that a connection was present when at least one streamline existed with endpoints in these two regions, resulting in a 90×90 connectivity matrix; NOS-weighted matrix (top), FA-weighted matrix (middle), SL-weighted matrix (bottom). AAL, automated anatomical labelling; FA, fractional anisotropy; NOS, number of streamlines; ROIs, regions of interest; SL, streamline length.

2.5 Network thresholding

The NOS-weighted matrix is typically used for thresholding (Sarwar et al., 2019; van Dellen et al., 2018). To obtain thresholded networks, edges with the lowest NOS are removed, which can result in disconnected networks. To ensure that the network did not become disconnected, we first calculated the minimum spanning tree (MST) using the Kruskal algorithm (Kruskal, 1956). The MST is an acyclic subgraph that connects all N nodes in the network such that the weight is minimal. We computed the MST by selecting the edges with the largest NOS first. This method of constructing the MST has previously been described as a good representation of the network backbone (Tewarie et al., 2015; van Dellen et al., 2018). The MST has per definition $N-1$ edges and a density of $2/N$. Since our network definition includes $N=90$ nodes, the density of the MST was $\approx 2.2\%$.

To obtain networks with larger densities, we added edges from the original unthresholded network to the MST, until a fixed density was reached, again by selecting edges with the largest NOS first. After each thresholding step, the NOS weights were disregarded to obtain binary matrices with densities of 2.2% (MST), 5%, 10%, 15%, and 20% (Figure 2). In this manner, the NOS weight determined which edges were retained after thresholding. The binary matrices were then multiplied with the FA-weighted and SL-weighted matrices for further analyses.

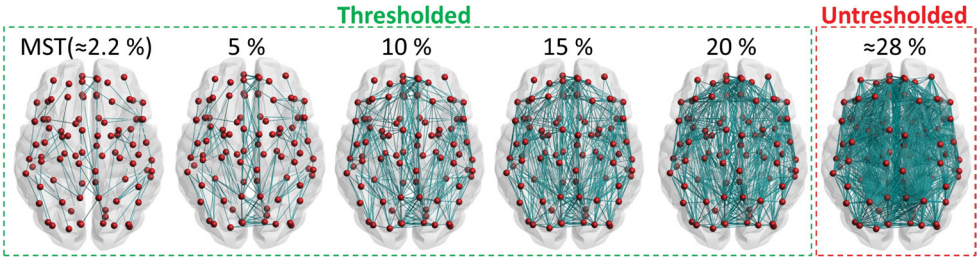


Figure 2: Network thresholding. The first step was the calculation of the MST (density ≈2.2%), by prioritizing the edges with the largest NOS. Next, edges were added until fixed densities of 5%, 10%, 15%, and 20% were reached. The average density of all the unthresholded networks of patients and controls was 28% ± 3% (mean ± SD). MST, minimum spanning tree; SD, standard deviation.

2.6 Edge composition and hubs

The first question we addressed was whether and how thresholding affects the characteristics of edges and nodes in the network. To examine the edge composition, we computed NOS, SL (mm), and FA of the edges removed versus the edges retained in the network.

We also explored whether the edges removed and retained during thresholding were preferentially connected to hub nodes. For this purpose, we first defined hub nodes as the nine nodes with the highest betweenness centrality (i.e., top 10th percentile) (Reijmer et al., 2015). Nodes with high betweenness centrality are often considered "hubs" because they participate in many shortest

paths and largely contribute to the global efficiency of the network (Hagmann et al., 2008; van den Heuvel and Sporns, 2011). After defining the hub nodes, we computed the percentage of hub connections relative to all edges retained in the network. Here, hub connections represent edges that connect hub nodes to other hub nodes and edges that connect hub nodes to non-hub nodes. We repeated the calculation of hub connections for each density level.

We also examined whether the hub nodes defined in the unthresholded network corresponded to the hub nodes defined in the thresholded networks. We computed nine hubs (according to the definition above) for each density level and compared their location with the location of the hubs in the unthresholded network. Finally, we investigated whether the overall nodal ranking (here defined as the ranked betweenness centrality of all nodes) was affected by thresholding. We calculated the nodal ranking for each density and compared it with the nodal ranking of the unthresholded network.

2.7 Detection of small vessel disease-related network impairments

The second question we addressed was whether thresholding influences comparisons of graph metrics between patients and controls. Ideally, the edge weight applied for thresholding the networks is not the same as applied for the between-group comparisons. In this study, we therefore obtained measures of global connectivity (e.g., global efficiency) and local connectivity (e.g., connectivity strengths of hub and non-hub nodes) from the FA-weighted networks. FA-weighted networks are frequently used to examine structural aspects of brain connectivity in patients and have been shown to be sensitive to the effects of SVD (Heinen et al., 2018; Lawrence et al., 2014; Reijmer et al., 2015; Tuladhar et al., 2016).

Global efficiency was defined as the inverse of the shortest path lengths (i.e., the minimum number of connections between each pair of brain regions). This network measure expresses the extent to which information is exchanged over the network (Rubinov and Sporns, 2010). Connectivity strength of hubs and non-hubs was defined as the mean FA over all the edges connected to hub nodes or non-hub nodes, respectively. All graph measures mentioned in this article were calculated using the brain connectivity toolbox (Rubinov and Sporns, 2010).

2.8 SVD imaging markers

SVD markers were computed as described previously (Ferro et al., 2019). Shortly, brain volumes were obtained using an in-house semiautomated pipeline as follows: (1) automated WMH segmentation using the FLAIR images; (2) lesion filling of 3D T1 images; and (3) probabilistic segmentations for gray matter, WM, and CSF. We visually checked the tissue classifications to ensure that the segmentations were correct and not affected by partial volume effects. Total brain volume was defined as the sum of the gray and WM volume. Brain volume was expressed as a percentage of the total intracranial volume. Lacunes were defined as hypointense areas between 2 and 15 mm on both FLAIR and T1-weighted images with a hyperintense rim on FLAIR images. Ratings were performed in line with the STRIVE (Standards for Reporting

Vascular Changes on Neuroimaging) criteria (Wardlaw et al., 2013) under supervision of a neuroradiologist (in training).

2.9 Statistical analysis

Statistical analysis was performed using IBM SPSS statistics (IBM Corp. Released 2017, IBM SPSS Statistics for Windows, Version 25.0.; IBM Corp., Armonk, NY). A p-value <0.05 was considered significant. Demographic variables and MRI markers of SVD were compared between patients and controls using independent-samples t-test for continuous variables and $\chi 2$-test for proportions. For the MMSE, brain volume, gray matter volume, and WMH volume, we used the nonparametric equivalent of the independent-samples t-test, Mann–Whitney U test, since these variables were not normally distributed. Characteristics of the unthresholded network (i.e., edge weights and network density) were compared between patients and controls using independent-samples t-test.

FA and SL were compared between the edges retained and the edges removed from the network using independent-samples t-test. Changes in the organization of nodes were analysed in two different ways: first, we computed the one-to-one overlap of hub nodes between the thresholded network and the thresholded networks; second, we correlated the ranked betweenness centrality of the thresholded network with the ranked betweenness centrality of the thresholded network using Pearson's correlation coefficient (r). Hence, the hub overlap reflects changes in the location of the nine hubs, while the correlation coefficient reflects changes in nodal ranking of all the nodes.

We analysed between-group differences (SVD vs. controls) in weighted graph metrics (global efficiency, strength of hub- and non-hub nodes) for each density level using a univariate ANOVA with age and sex as covariates. Effect sizes were defined as Cohen's d, and p-values were adjusted for multiple testing using Bonferroni correction. To examine whether the between-group differences change across density levels, we tested interactions between density × group using a repeated-measures ANOVA with group as between-subject factor and density as within-subject factor. Again, age and sex were used as covariates.

3 RESULTS

3. 1 Characteristics of the unthresholded network

Characteristics of the unthresholded networks of patients and controls are shown in Figure 3. The NOS of all edges in the unthresholded network was similar for patients and controls ($p > 0.14$; Figure 3A). However, the unthresholded networks of patients were composed of edges with relatively long streamlines ($p < 0.001$; Figure 3B) and lower FA ($p < 0.001$; Figure 3C). Of note, the edge FA and edge length seemed independent of the NOS (Supplementary Figure S1). The network density of the unthresholded networks of patients (mean ± SD = 26.98% ± 3.12%) was lower when compared with the unthresholded networks of controls (mean ± SD = 28.78% ± 3.17%) ($p = 0.002$, Figure 3D).

Nodes that were identified as hubs in the unthresholded networks of patients and controls are shown in Figure 3E and F. Hub nodes were largely consistent across groups and with those commonly reported in the literature (Hagmann et al., 2008; van den Heuvel and Sporns, 2011, 2013).

3.2 Effect of thresholding on edge composition and hubs

The effect of thresholding on the edge composition of all subjects is shown in Figure 4. The edges removed from the network during thresholding had, by definition, a lower NOS, but they were also on average longer than the edges retained ($p < 0.001$; Figure 4A, B). However, the mean FA was similar for both sets of removed edges ($p > 0.10$; Figure 4C).

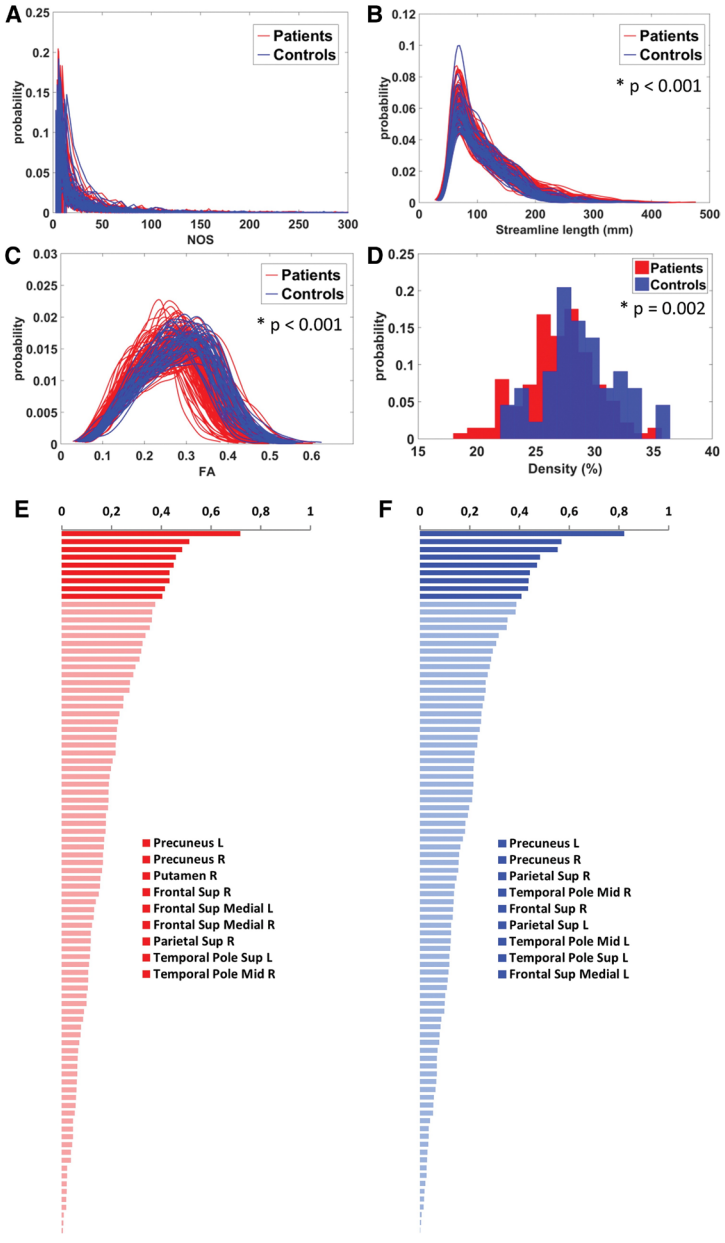


Figure 3: Characteristics of the unthresholded networks. A: NOS weight of all edges in the unthresholded networks of patients (red) and controls (blue). B: SL of all edges in the unthresholded networks (patients vs. controls $p<0.001$). C: FA weight of all edges in the unthresholded networks (patients vs. controls $p<0.001$). D: Histogram of network density of the unthresholded networks of patients (mean ± SD = 26.98% ± 3.12%) and controls (mean ± SD = 28.78% ± 3.17%), $p=0.002$. E: Average node betweenness centrality of the unthresholded networks of patients. F: Average node betweenness of the unthresholded networks of controls. Hubs were defined as the nine nodes with the highest betweenness.

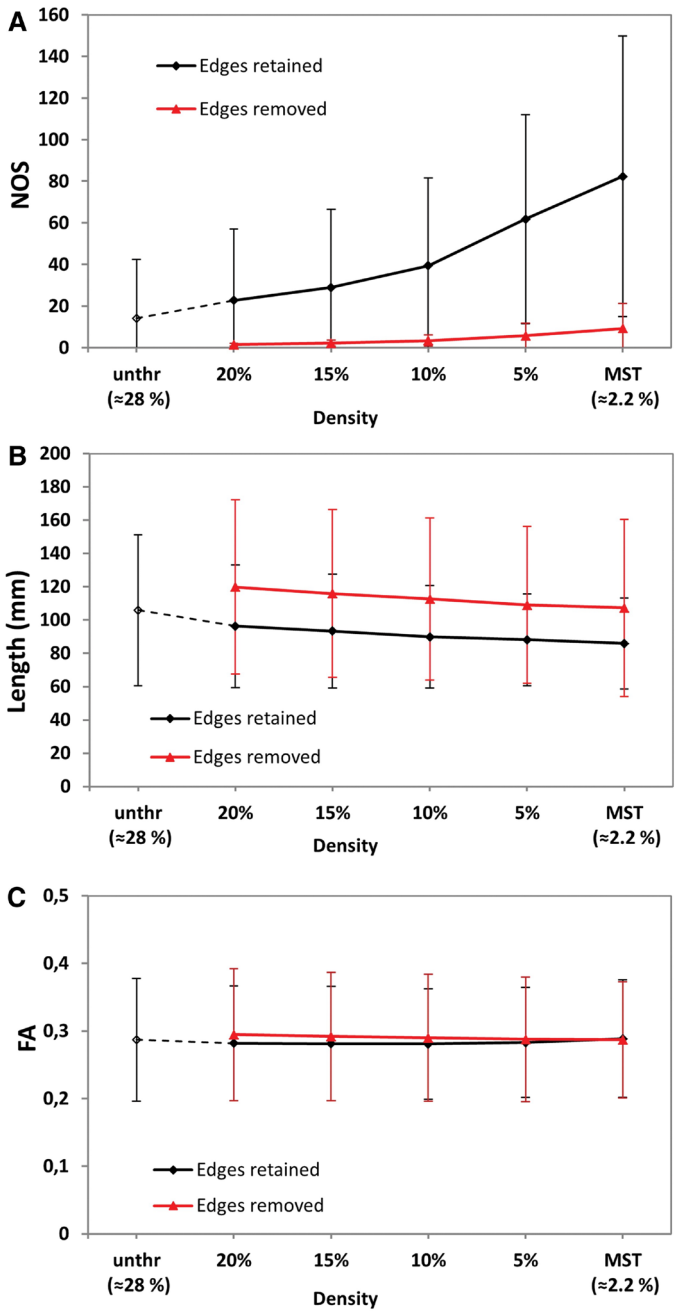


Figure 4: Effect of thresholding on the edge composition. A: Average NOS of edges retained (black) and edges removed (red) at each density level (all $p<0.001$). Error bars represent the SD. "unthr" (unthresholded network). B: Average SL of the edges retained, and edges removed at each density level (all $p<0.001$); C: average FA of the edges retained, and the edges removed at each density level (all $p>0.1$). p-Values were derived from independent-samples t-tests.

The effect of thresholding on hub connections and hub nodes of all subjects is shown in Figure 5. In the unthresholded network, 33% of all edges were hub connections (Figure 5A). This proportion did not change after thresholding, indicating that thresholding did not preferentially target edges connected to hub or non-hub nodes. These results indicate that fixed-density thresholding (based on NOS weights) removes edges composed of relatively long streamlines but is independent of FA and does not disproportionally affect hub connections. When a small proportion of edges was removed, more than half of the hubs found in the thresholded network corresponded with the original hubs found in the unthresholded network (hub overlap ≥5/9 for densities ≥10%; Figure 5B). When a greater proportion of edges was removed, the majority of hubs defined in the unthresholded network no longer corresponded with the hubs defined in the thresholded network (hub overlap ≤ 4/9 for densities ≤5%, Figure 5B).

We found similar results when we compared the overall nodal ranking (i.e., the ranked betweenness centrality of all nodes) (Figure 5C). There was a high correspondence in nodal ranking when a small proportion of edges was removed ($r > 0.77$ for densities ≥10%) and lower correspondence in nodal ranking when a larger proportion of edges was removed ($r < 0.64$ for densities ≤5%).

When we performed these analyses separately for patients and controls, fixed-density thresholding had a similar effect on the edge composition, hub location, and hub connections for both groups (Supplementary Figures S2 and S3).

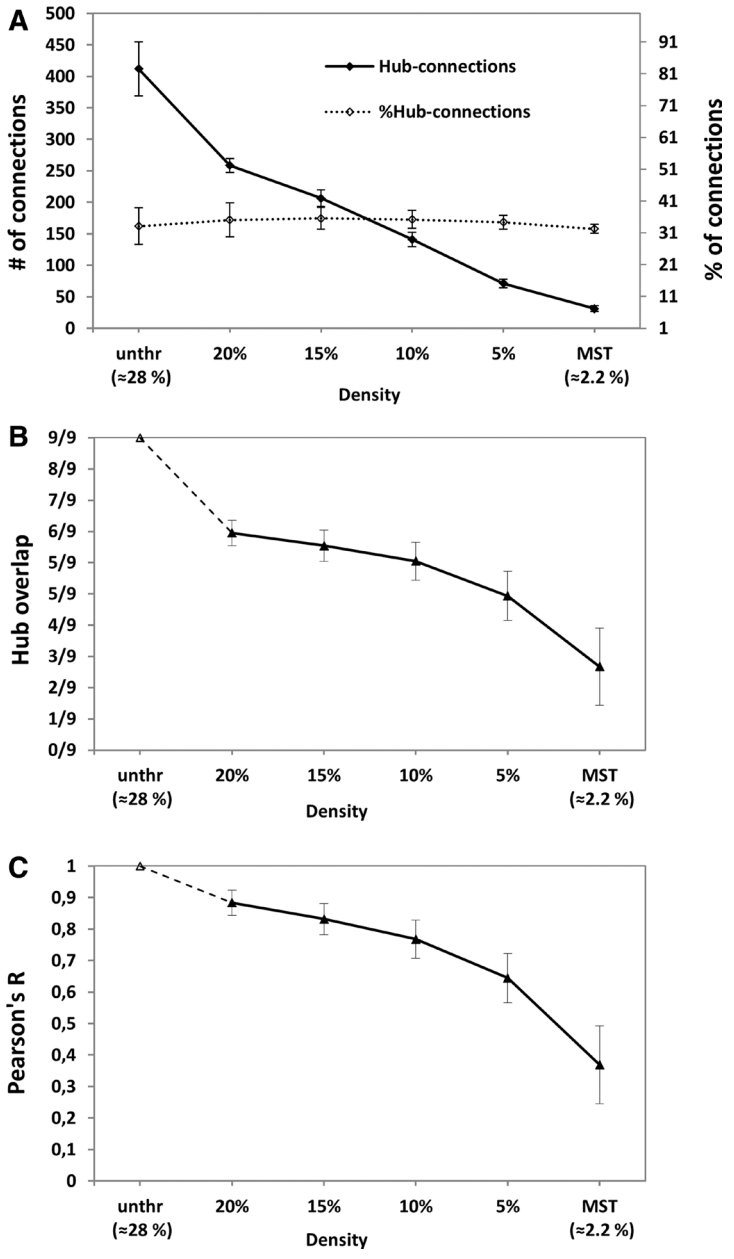


Figure 5: Effect of thresholding on hub connections and hub nodes. A: Average number of hub connections retained in the network at each density level. This number is also represented as the percentage of all the edges retained in the network (dotted line, right y-axis). Error bars represent the SD. B: Overlap between the hubs defined in the unthresholded network and the hubs defined in the thresholded networks (20%, 15%, 10%, 5% and MST). C: Correlation between the nodal ranking of the unthresholded and nodal ranking of the thresholded networks (20%, 15%, 10%, 5% and MST). Note that each density is compared against the unthresholded network. Thus, the hub overlap and correlation coefficient for the unthresholded network are 9/9 and 1, respectively.

3.3 Effect of thresholding on the detection of SVD-related network impairments

In line with the previous literature (Heinen et al., 2018; Lawrence et al., 2014; Tuladhar et al., 2016), the global efficiency of the unthresholded FA-weighted network was lower in patients with SVD when compared with controls (Cohen's $d = 0.97$, $p < 0.001$; Figure 6A). The between-group differences in global efficiency remained present after thresholding, irrespective of the network density (Figure 6A). The largest effect size was observed at 15% density (Cohen's $d = 1.09$, interaction term density × group: $p = 0.026$).

The connectivity strength of hub nodes in the unthresholded network was lower in patients when compared with controls (Cohen's $d = 1.11$, $p < 0.001$, Figure 6B). Again, the between-group difference remained significant after thresholding. The largest effect size in strength of hub nodes was observed at 10% density (Cohen's $d = 1.18$, interaction term density × group: $p < 0.001$).

Similarly, the largest effect size in strength of non-hub nodes was observed at 15% density (Cohen's $d = 0.95$, interaction term density × group: $p < 0.001$, Figure 6C). At 10% and 5% densities, the group difference in strength of non-hub nodes was not significant and even reversed, indicating that the strength of non-hub nodes is less stable over thresholds.

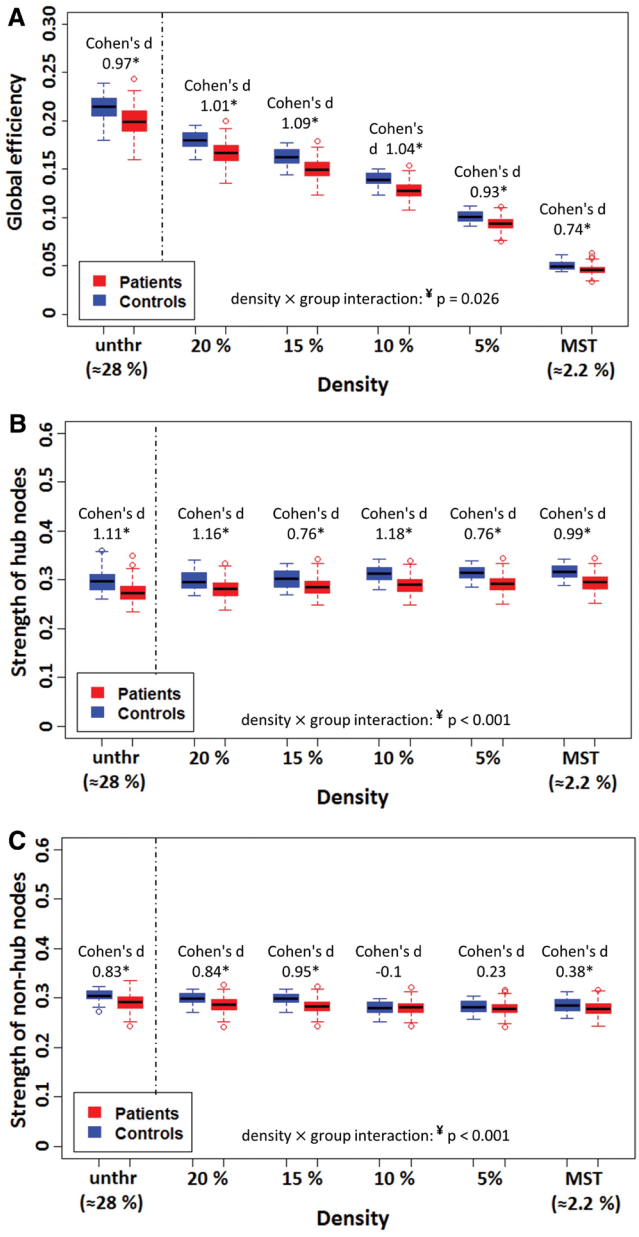


Figure 6: Difference in weighted network measures between patients with SVD (red) and controls (blue) across different levels of network density. A: FA-weighted global network efficiency. B: Average FA-weighted strength of hub nodes. C: Average FA-weighted strength of non-hub nodes. Hubs were defined in the control group; thus, the same nodes were considered hubs in both groups. *p<0.001, Bonferroni corrected; age- and sex-adjusted effect sizes are given as Cohen's d. ¥p≤0.026, density×group interaction. SVD, small vessel disease.

4 DISCUSSION

In this study, we investigated the impact of fixed-density thresholding on the basic architecture of structural brain networks and on the subsequent detection of network abnormalities in patients with SVD. Our results showed that fixed-density thresholding mainly removes edges composed of long streamlines but seemed independent of FA. The removed edges were not preferentially connected to hub or to non-hub nodes. As a result, more than half of the hubs and their original location, as well as the overall nodal ranking, were preserved when networks were thresholded to a density >10%. Importantly, our results also showed that the sensitivity to detect network abnormalities in patients with SVD is not affected by thresholding. Patients had a lower global efficiency and hub strength than controls when we compared their unthresholded networks. These between-group differences remained present in the thresholded networks, indicating that the disease effects are not removed by thresholding. Thresholding of brain networks in case-control studies remains a topic of intense debate (Stam, 2014). Some have argued that the difference in network density is partly a consequence of the disease and enforcing a similar density would therefore lead to an elimination of disease effects. Here, we showed that disease effects were not removed by obtaining measures from brain networks with fixed densities.

Another argument against fixed-density thresholding in clinical network studies is that group differences may be artificially inflated. This effect is especially problematic when there is an initial group difference in the edge weights used for thresholding (van den Heuvel et al., 2017). Under these circumstances, more high-weight edges will be selected from the control networks than the patient networks to obtain the same network density, thereby increasing the between-group difference in network weight. This inflation of between-group differences is often a problem in functional network studies where the same measure of functional connectivity is used for thresholding as well as for calculation of network measures for group comparison. In our study however, the unthresholded networks of patients and controls had similar NOS weight distributions. Furthermore, the edge weight used to calculate graph metrics (i.e., FA) seemed independent of the weight used for thresholding (i.e., NOS). Therefore, this form of bias introduced by the difference in thresholding weight was not an issue.

We also found that a larger proportion of long-range connections was removed during thresholding, indicating that longer connections are often composed of low NOS (Roberts et al., 2017). A possible explanation for this bias is that long-range white matter pathways are inherently more difficult to tract because the likelihood of meeting the stopping criteria of the tractography algorithm (i.e., extreme fibre deviations or voxels with low FOD) is higher for long-range fibres (Jeurissen et al., 2017; Leopold et al., 2014). Understanding how the connection length

is affected by thresholding is relevant because the role of long-range connections is a topic of interest in disease studies (Tuladhar et al., 2017), due to their contribution for global network signalling (van den Heuvel et al., 2012).

The edges removed during thresholding were not preferentially connected to hub nodes (or to non-hub nodes). In this way, the removal of edges seemed comparable with a "random network attack" (Mengiste et al., 2015). Brain networks are shown to be resilient to random attacks due to their hierarchical organization (Crossley et al., 2014). This resilience could also explain why the hub location and nodal ranking were largely unaffected when the networks were thresholded to densities ≥10%. These results suggest that a threshold between 10% and 15% can be applied without significantly damaging the backbone structure of brain networks. This is also supported by previous research that shows that brain networks preserve their small-world characteristics when they are thresholded to densities between 10% and 34% (Zhang et al., 2011).

Finally, we showed that the between-group differences in graph metrics are preserved after thresholding. In line with previous work, global efficiency and FA-weighted strength of hubs in the unthresholded network were lower in patients with SVD than in controls (Lawrence et al., 2014; Reijmer et al., 2015). These group differences remained present after thresholding, which could be explained by the fact that there was no correlation between the NOS of the edges removed and their FA value. Thus, the FA weight used to calculate weighted graph metrics was not influenced by thresholding. Again, this shows the importance of using independent edge weights for thresholding and for the calculation of weighted graph metrics. Another explanation for the preservation of group differences is that thresholding had a similar impact on the network architecture of patients and controls. Thus, the computation of graph metrics was not affected differently for both groups. The only measure where the group differences were not maintained (for densities <10%) was the strength of non-hub nodes. Since these nodes have low degree to begin with, when the networks are thresholded to low densities, local graph metrics are calculated based on few connections (often only one), making the strength of non-hub nodes a less stable measure.

The fact that the largest effect sizes were observed at 10% and 15% densities suggests that fixed-density thresholding might improve the detection of between-group differences. It is possible that by applying moderate thresholds, the amount of spurious (false-positive) connections is reduced without significantly increasing the number of false negatives (i.e., true connections removed during thresholding), which can improve the network specificity and sensitivity. Indeed, previous research indicates that reducing the number of false positives in the network is more beneficial than safeguarding against the creation of new false negatives (Zalesky et al., 2016). These mid-range densities (10% and 15%) could represent the levels where this balance between

false positives and false negatives is achieved. Thresholding beyond these levels would be detrimental to both the initial network architecture and the detection of between-group differences.

Strength of this study includes a standardized clinical MRI protocol with high-quality diffusion data from the TRACE-VCI cohort (Boomsma et al., 2017). Furthermore, we performed whole-brain tractography using CSD, which is known to have a better anatomical accuracy than the conventional diffusion tensor imaging approach (Jeurissen et al., 2013). Moreover, by incorporating the MST into the thresholded network, we ensured that the network did not become disconnected, avoiding differences in network structure due to disconnected nodes.

This study also has some limitations. Brain network reconstruction involves several choices in pre-processing steps, which could have influenced our results. The choice of parcellation scheme (Zalesky et al., 2010), tractography algorithm (Bastiani et al., 2012), and tractography parameters (e.g., FOD threshold, l-max, streamline deflection angle, etc.) can have a significant impact on the definition of nodes and edges as well as on the number of false positives and false negatives in the network. Consequently, these choices can influence the thresholding results and graph outcome measures obtained from these networks. Another limitation is that we did not consider the effect of thresholding on all network characteristics but only on network properties commonly investigated in network studies with SVD populations. Furthermore, we used the betweenness centrality as the only metric to characterize nodal ranking and define hub nodes. This may not fully describe the importance of a node for network integration and communication. Alternatively, a metric such as communicability (Andreotti et al., 2014a) could have provided a broader view of node importance. However, the interpretation of this metric is not straightforward and not established in clinical studies.

Future studies should examine the consequences of fixed-density thresholding on other network parameters and in other disease populations. Nevertheless, fixed-density thresholding may yield more reproducible networks, as suggested by repeatability studies (Andreotti et al., 2014b; Roine et al., 2018; Tsai et al., 2018). This can be beneficial in studies with a longitudinal design, but further investigation is required.

5 CONCLUSIONS

We showed that moderate fixed-density thresholds (10% and 15% densities in this dataset) can be used to control for density effects in diffusion-based brain network studies. The present findings can help researchers make a well-informed choice when selecting an appropriate density threshold for future clinical studies.

ACKNOWLEDGMENTS

Members of the Utrecht Vascular Cognitive Impairment (VCI) study group involved in this study (in alphabetical order by department): University Medical Center Utrecht, The Netherlands, Department of Neurology: E. van den Berg, J.M. Biesbroek, G.J. Biessels, M. Brundel, W.H. Bouvy, L.G. Exalto, C.J.M. Frijns, O. Groeneveld, S.M. Heringa, N. Kalsbeek, L.J. Kappelle, Y.D. Reijmer, J. Verwer; Department of Radiology/Image Sciences Institute: J. de Bresser, H.J. Kuijf, A. Leemans, P.R. Luijten, M.A. Viergever, K.L. Vincken, J.J.M. Zwanenburg; Department of Geriatrics: H.L. Koek; Hospital Diakonessenhuis Zeist, The Netherlands: M. Hamaker, R. Faaij, M. Pleizier, E. Vriens.

SUPPLEMENTARY INFORMATION

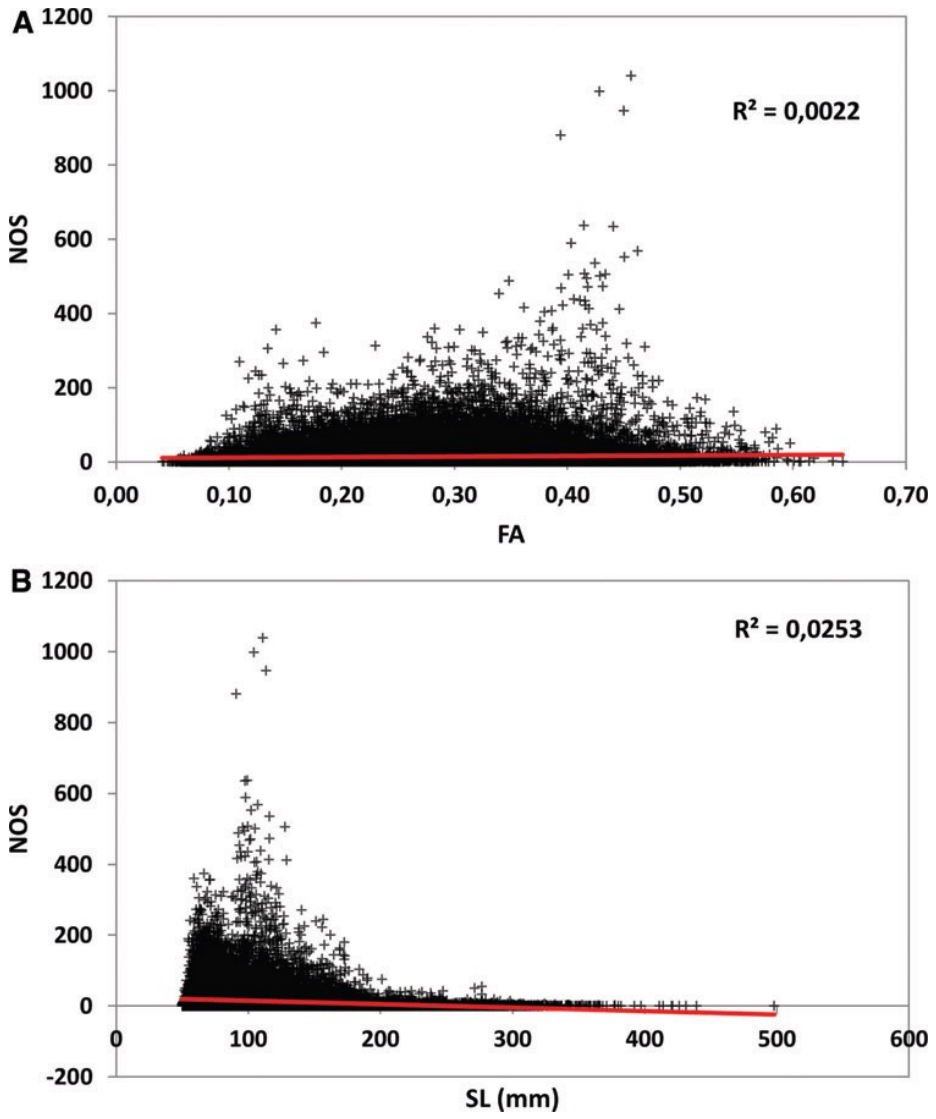


Figure S1: Data points represent the composition of all edges found in the unthresholded networks of all patients and all controls. Red lines represent the linear regression fit between FA and NOS. A: Correlation between the NOS and edge FA. Edges with the highest NOS did not necessarily have the highest FA (R2 = 0.0022). B: Correlation between NOS and average SL. Despite the low correlation (R2 = 0.0253), the longest edges (>300 mm) are composed of few streamlines and therefore are removed first during thresholding. FA, fractional anisotropy; NOS, number of streamlines; SL, streamline length.

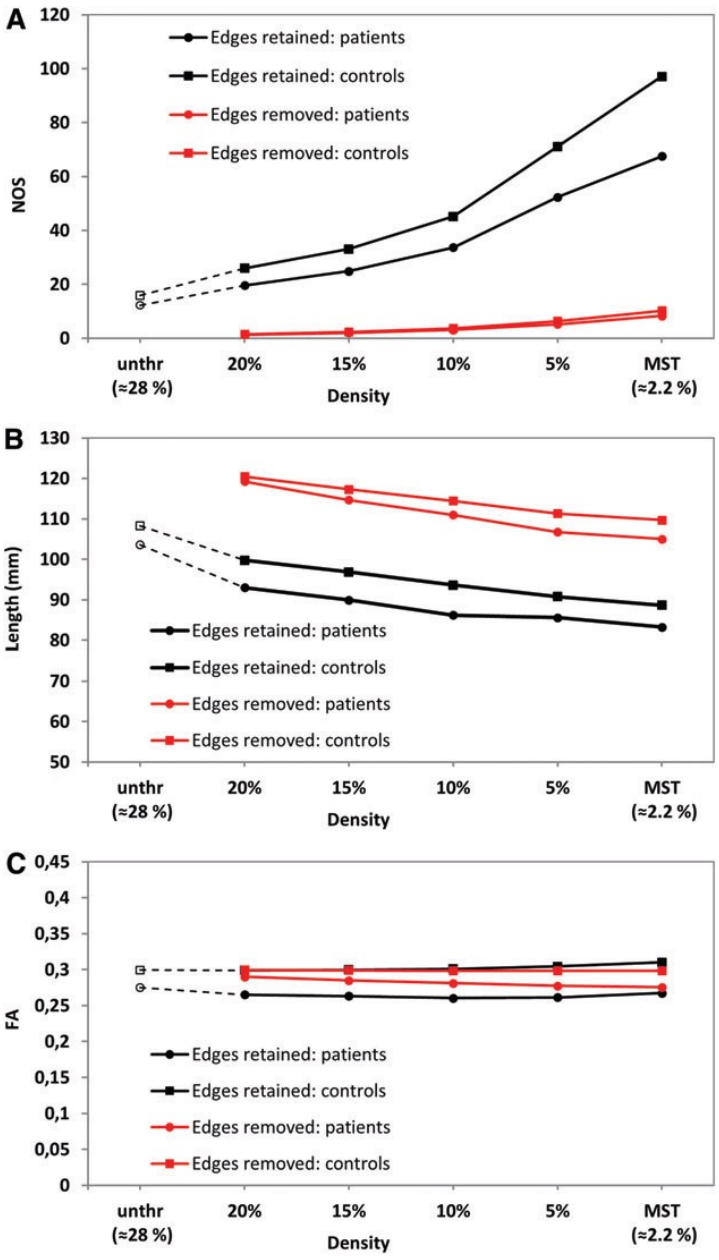


Figure S2: Effect of thresholding on the edge composition. Patients and controls are shown separately. A: Average NOS of edges retained (black) and edges removed (red) at each density level. Error bars are omitted for simplification of the plot. ''unthr'' (unthresholded network). B: Average streamline length of the edges retained, and edges removed at each density level. C: Average FA of the edges retained, and the edges removed at each density level.

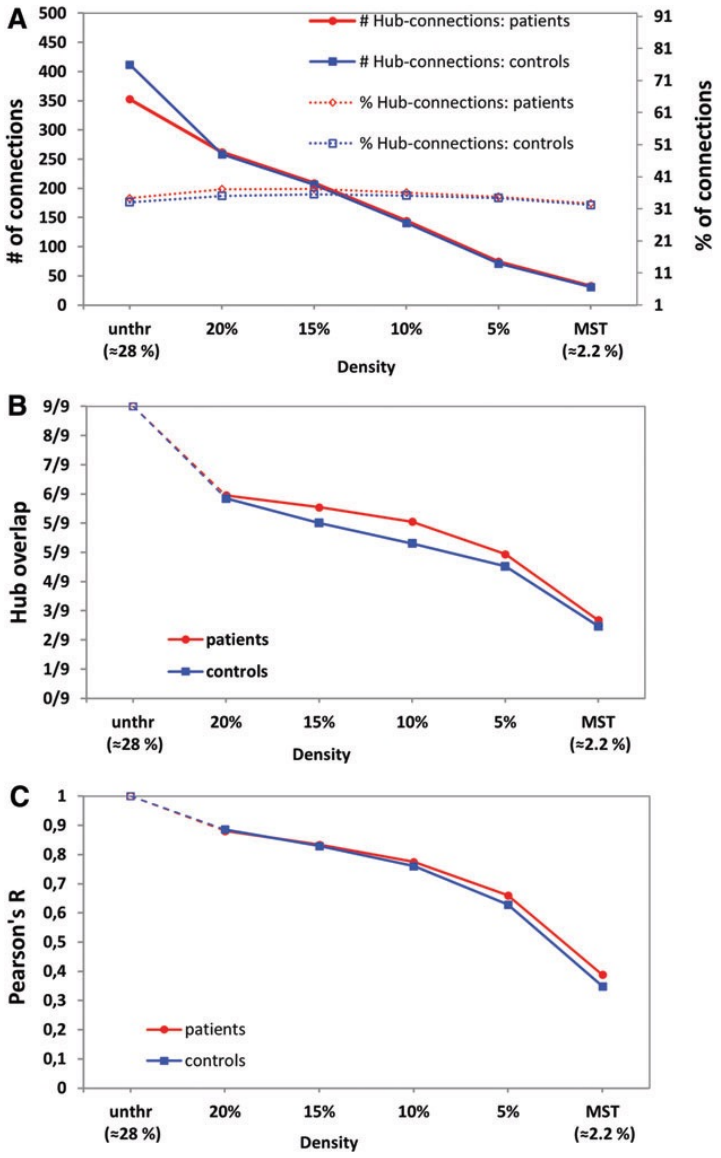


Figure S3: Effect of thresholding on hub connections and hub nodes. Patients (red) and controls (blue) are shown separately. A: Average number of hub connections retained in the network at each density level. This number is also represented as the percentage of all the edges retained in the network (dotted line, right y-axis). Error bars are omitted for simplification of the plot. ''unthr'' (unthresholded network). B: Overlap between the hubs defined in the unthresholded network and the hubs defined in the thresholded networks (20%, 15%, 10%, 5% and MST). C: Correlation between the nodal ranking of the unthresholded and nodal ranking of the thresholded networks (20%, 15%, 10%, 5% and MST). Note that each density is compared against the unthresholded network. Thus, the hub overlap and correlation coefficient for the unthresholded network are 9/9 and 1, respectively. MST, minimum spanning tree.

REFERENCES

Achard S, Bullmore E. 2007. Efficiency and cost of economical brain functional networks. PLoS Comput Biol 3:0174–0183. Crossref.

Achard S, Delon-Martin C, Vertes PE, Renard F, Schenck M, Schneider F, et al. 2012. Hubs of brain functional networks are radically reorganized in comatose patients. Proc Natl Acad Sci U S A 109:20608–20613.

Andreotti J, Abela E, Dierks T, Federspiel A, Melie-Garcia L, Jann K, et al. 2014a. Validation of network communicability metrics for the analysis of brain structural networks. PLoS One 9:e115503.

Andreotti J, Jann K, Melie-Garcia L, Dierks T, Federspiel A, Giezendanner S. 2014b. Repeatability analysis of global and local metrics of brain structural networks. Brain Connect 4:203–220.

Bassett DS, Bullmore ET. 2009. Human brain networks in health. Curr Opin Neurol 10:324–336.

Bassett DS, Bullmore ET, Meyer-lindenberg A, Coppola R. 2009. Cognitive fitness of cost-efficient brain functional networks. Proc Natl Acad Sci U S A 106:11747–11752.

Bastiani M, Shah NJ, Goebel R, Roebroeck A. 2012. Human cortical connectome reconstruction from diffusion weighted MRI: the effect of tractography algorithm. Neuroimage 62:1732–1749.

Boomsma JMF, Exalto LG, Barkhof F, van den Berg E, de Bresser J, Heinen R, et al. 2017. Vascular cognitive impairment in a memory clinic population: rationale and design of the "Utrecht-Amsterdam Clinical Features and Prognosis in Vascular Cognitive Impairment" (TRACE-VCI) study. JMIR Res Protoc 6:e60.

Crossley NA, Mechelli A, Scott J, Carletti F, Fox PT, Mcguire P, Bullmore ET. 2014. The hubs of the human connectome are generally implicated in the anatomy of brain disorders. Brain 137:2382–2395.

Daianu M, Jahanshad N, Nir TM, Jack CR, Weiner MW, Bernstein MA, Thompson PM. 2015. Rich club analysis in the Alzheimer's disease connectome reveals a relatively undisturbed structural core network. Hum Brain Mapp 36:3087–3103.

Daianu M, Mezher A, Mendez MF, Jahanshad N, Jimenez EE, Thompson PM. 2016. Disrupted rich club network in behavioral variant frontotemporal dementia and early-onset Alzheimer's disease. Hum Brain Mapp 37:868–883.

Drakesmith M, Caeyenberghs K, Dutt A, Lewis G, David AS, Jones DK. 2015. Overcoming the effects of false positives and threshold bias in graph theoretical analyses of neuroimaging data. Neuroimage 118:313–333.

Fagerholm ED, Hellyer PJ, Scott G, Leech R, Sharp DJ. 2015. Disconnection of network hubs and cognitive impairment after traumatic brain injury. Brain 138:1696–1709.

Ferro D, Heinen R, De Brito Robalo B, Kuijf H, Biessels GJ, Reijmer Y. 2019. Cortical microinfarcts and white matter connectivity in memory clinic patients. Front Neurol 10:571.

Garrison KA, Scheinost D, Finn ES, Shen X, Constable RT. 2015. The (in)stability of functional brain network measures across thresholds. Neuroimage 118:651–661.

Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey CJ, Van Wedeen J, Sporns O. 2008. Mapping the structural core of human cerebral cortex. PLoS Biol 6:1479–1493. Crossref.

Hein KHM, Neher P, Christophe J, Alexandre M. 2016. Tractography-based connectomes are dominated by false-positive connections. BioRxiv 1–23. Crossref.

Heinen R, Vlegels N, de Bresser J, Leemans A, Biessels GJ, Reijmer YD. 2018. The cumulative effect of small vessel disease lesions is reflected in structural brain networks of memory clinic patients. Neuroimage 19:963–969.

Jeurissen B, Descoteaux M, Mori S, Leemans A. 2017. Diffusion MRI fiber tractography of the brain. NMR Biomed 32:e3785.

Jeurissen B, Leemans A, Jones DK, Tournier JD, Sijbers J. 2011. Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. Hum Brain Mapp 32:461–479.

Jeurissen B, Leemans A, Tournier JD, Jones DK, Sijbers J. 2013. Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging. Hum Brain Mapp 34:2747–2766.

Jones DK. 2008. Studying connections in the living human brain with diffusion MRI. Cortex 44:936–952.

Kruskal JB. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc Am Math Soc 7:48. Crossref.

Lawrence AJ, Chung AW, Morris RG, Markus HS, Barrick TR. 2014. Structural network efficiency is associated with cognitive impairment in small-vessel disease. Neurology 83:304–311.

Leemans A, Jeruissen B, Sijbers J, Jones DK. 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. Proc Int Soc Magn Reson Med 245:3537.

Leemans A, Jones DK. 2009. The B-matrix must be rotated when correcting for subject motion in DTI data. Magn Reson Med 61:1336–1349.

Leopold DA, Modi P, Thomas C, Irfanoglu MO, Saleem KS, Ye FQ, Pierpaoli C. 2014. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. Proc Natl Acad Sci U S A 111:16574–16579.

Liu J, Li M, Pan Y, Lan W, Zheng R, Wu FX, Wang J. 2017. Complex brain network analysis and its applications to brain disorders: a survey. Complex 2017:1–27. Crossref. Crossref.

Mårtensson G, Pereira JB, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, et al. 2018. Stability of graph theoretical measures in structural brain networks in Alzheimer's disease. Sci Rep 8:1–15.

Mengiste SA, Aertsen A, Kumar A. 2015. Effect of edge pruning on structural controllability and observability of complex networks. Sci Rep 5:1–14. Crossref.

Odish OFF, Caeyenberghs K, Hosseini H, Van Den Bogaard SJA, Roos RAC, Leemans A. 2015. Dynamics of the connectome in Huntington's disease: a longitudinal diffusion MRI study. Neuroimage 9:32–43.

Pandit AS, Expert P, Lambiotte R, Bonnelle V, Leech R, Turkheimer FE, Sharp DJ. 2013. Traumatic brain injury impairs small-world topology. Neurology 80:1826–1833.

Qi S, Meesters S, Nicolay K, ter Haar Romeny BM, Ossenblok P. 2015. The influence of construction methodology on structural brain network measures: a review. J Neurosci Methods 253:170–182.

Reijmer YD, Fotiadis P, Martinez-Ramirez S, Salat DH, Schultz A, Shoamanesh A, et al. 2015. Structural network alterations and neurological dysfunction in cerebral amyloid angiopathy. Brain 138:179–188.

Roberts JA, Perry A, Roberts G, Mitchell PB, Breakspear M. 2017. Consistency-based thresholding of the human connectome. Neuroimage 145:118–129.

Roine T, Aelterman J, Sijbers J, Jeurissen B, Philips W, Leemans A, Perrone D. 2018. Reproducibility and intercorrelation of graph theoretical measures in structural brain connectivity networks. Med Image Anal 52:56–67.

Rubinov M, Sporns O. 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52:1059–1069.

Sarwar T, Ramamohanarao K, Zalesky A. 2019. Mapping connectomes with diffusion MRI: deterministic or probabilistic tractography? Magn Reson Med 81:1368–1384.

Sporns O, Tononi G, Kötter R. 2005. The human connectome: a structural description of the human brain. PLoS Comput Biol 1:0245–0251. Crossref.

Stam CJ. 2014. Modern network science of neurological disorders. Nat Rev Neurosci 15:683–695.

Tax CMW, Otte WM, Viergever MA, Dijkhuizen RM, Leemans A. 2015. REKINDLE: Robust Extraction of Kurtosis INDices with linear estimation. Magn Reson Med 73:794–808.

Telesford QK, Simpson SL, Burdette JH, Hayasaka S, Laurienti PJ. 2011. The brain as a complex system: using network science as a tool for understanding the brain. Brain Connect 1:295–308.

Tewarie P, van Dellen E, Hillebrand A, Stam C.J. 2015. The minimum spanning tree: an unbiased method for brain network analysis. Neuroimage 104:177–188.

Tsai SY. 2018. Reproducibility of structural brain connectivity and network metrics using probabilistic diffusion tractography. Sci Rep 8:1–12.

Tuladhar AM, Lawrence A, Norris DG, Barrick TR, Markus HS, de Leeuw FE. 2017. Disruption of rich club organisation in cerebral small vessel disease. Hum Brain Mapp 38:1751–1766.

Tuladhar AM, van Dijk E, Zwiers MP, van Norden AGW, de Laat KF, Shumskaya E, et al. 2016. Structural network connectivity and cognition in cerebral small vessel disease. Hum Brain Mapp 37:300–310.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–289.

van Dellen E, Sommer IE, Bohlken MM, Tewarie P, Draaisma L, Zalesky A, et al. 2018. Minimum spanning tree analysis of the human connectome. Hum Brain Mapp 39:2455–2471.

Van Den Berg E, Reijmer YD, De Bresser J, Kessels RPC, Kappelle LJ, Biessels GJ. 2010. A 4 year follow-up study of cognitive functioning in patients with type 2 diabetes mellitus. Diabetologia 53:58–65.

van den Heuvel MP, de Lange SC, Zalesky A, Seguin C, Yeo BTT, Schmidt R. 2017. Propor-

tional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: issues and recommendations. Neuroimage 152:437–449.

van den Heuvel MP, Kahn RS, Goni J, Sporns O. 2012. High-cost, high-capacity backbone for global brain communication. Proc Natl Acad Sci U S A 109:11372–11377.

van den Heuvel MP, Sporns O. 2011. Rich-club organization of the human connectome. J Neurosci 31:15775–15786.

van den Heuvel MP, Sporns O. 2013. Network hubs in the human brain. Trends Cogn Sci 17:683–696.

van den Heuvel MP, Stam CJ, Boersma M, Hulshoff Pol HE. 2008. Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. Neuroimage 15:528–539. Crossref.

van Wijk BCM, Stam CJ, Daffertshofer A. 2010. Comparing brain networks of different size and connectivity density using graph theory. PLoS One 5: 13701. Crossref.

Váša F, Bullmore ET, Patel AX. 2018. Probabilistic thresholding of functional connectomes: application to schizophrenia. Neuroimage 172:326–340.

Veraart J, Sijbers J, Sunaert S, Leemans A, Jeurissen B. 2013. Weighted linear least squares estimation of diffusion MRI parameters: strengths, limitations, and pitfalls. Neuroimage 81:335–346.

Vos SB, Tax CMW, Luijten PR, Ourselin S, Leemans A, Froeling M. 2017. The importance of correcting for signal drift in diffusion MRI. Magn Reson Med 77:285–299.

Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, et al. 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol 12:822–838.

Zalesky A, Fornito A, Cocchi L, Gollo LL, van den Heuvel MP, Breakspear M. 2016. Connectome sensitivity or specificity: which is more important? Neuroimage 142:407–420.

Zalesky A, Fornito A, Harding IH, Cocchi L, Yücel M, Pantelis C, Bullmore ET. 2010. Whole-brain anatomical networks: does the choice of nodes matter? Neuroimage 50:970–983.

Zhang J, Wang J, Wu Q, Kuang W, Huang X, He Y, Gong Q. 2011. Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder. Biol Psychiatry 70:334–342.

Chapter 4

Impact of thresholding on the consistency and sensitivity of diffusion MRI-based brain networks in patients with cerebral small vessel disease

Bruno M. de Brito Robalo
Naomi Vlegels
Alexander Leemans
Yael D. Reijmer
Geert Jan Biessels

Brain and Behavior 2022 Apr, e2523
DOI: 10.1002/brb3.2523

ABSTRACT

Introduction: Thresholding of low-weight connections of diffusion MRI-based brain networks has been proposed to remove false-positive connections. It has been previously established that this yields more reproducible scan–rescan network architecture in healthy subjects. In patients with brain disease, network measures are applied to assess inter-individual variation and changes over time. Our aim was to investigate whether thresholding also achieves improved consistency in network architecture in patients, while maintaining sensitivity to disease effects for these applications.

Methods: We applied fixed-density and absolute thresholding on brain networks in patients with cerebral small vessel disease (SVD, n = 86; ≈ 24 months follow-up), as a clinically relevant exemplar condition. In parallel, we applied the same methods in healthy young subjects (n = 44; scan–rescan interval ≈4 months) as a frame of reference. Consistency of network architecture was assessed with dice similarity of edges and intraclass correlation coefficient (ICC) of edge-weights and hub-scores. Sensitivity to disease effects in patients was assessed by evaluating interindividual variation, changes over time, and differences between those with high and low white matter hyperintensity burden, using correlation analyses and mixed ANOVA.

Results: Compared to unthresholded networks, both thresholding methods generated more consistent architecture over time in patients (unthresholded: dice = 0.70; ICC: 0.70–0.78; thresholded: dice = 0.77; ICC: 0.73–0.83). However, absolute thresholding created fragmented nodes. Similar observations were made in the reference group. Regarding sensitivity to disease effects in patients, fixed-density thresholds that were optimal in terms of consistency (densities: 0.10–0.30) preserved interindividual variation in global efficiency and node strength as well as the sensitivity to detect effects of time and group. Absolute thresholding produced larger fluctuations of interindividual variation.

Conclusions: Our results indicate that thresholding of low-weight connections, particularly when using fixed-density thresholding, results in more consistent network architecture in patients with longer rescan intervals, while preserving sensitivity to disease effects.

1 INTRODUCTION

Diffusion-weighted imaging (DWI) and fibre tractography enable us to map cerebral white matter pathways and reconstruct large-scale brain networks (Jeurissen et al., 2017). Subsequently, graph theory can be applied to quantify the properties of such networks (Hagmann et al., 2007; Sporns et al., 2005). This framework has been widely used to investigate not only normal brain development but also a variety of neurological and psychiatric disorders (Fornito et al., 2013; Tijms et al., 2013).

A major challenge in structural network analysis is the limited reproducibility of networks obtained with diffusion MRI, due to the presence of false-negative and false-positive connections (Buchanan et al., 2014; Zalesky et al., 2016). Essentially, false negatives represent white matter connections that are undetected by the tractography algorithm. By contrast, false positives are edges in the reconstructed network that do not represent true white matter connections. These errors can result from numerous processing steps (Sotiropoulos & Zalesky, 2017), particularly the choice of parcellation scheme (Zalesky et al., 2010), tractography algorithm (Sarwar et al., 2018), and weighting strategy (Dimitriadis et al., 2017).

The most common solution to reduce false positives is to employ weight-based thresholding by removing so-called "weak" connections. In diffusion-based brain networks, weak connections are usually defined as having a low number of streamlines (NOS). Many network studies have used thresholding strategies such as absolute thresholding (Garrison et al., 2015; Nicols et al., 2016), where a uniform threshold is applied to remove all connections below a certain edge-weight (e.g., below five streamlines). Another popular method is fixed-density thresholding (Rubinov & Sporns, 2010; van den Heuvel et al., 2017), where a relative threshold derived from each individual's data is applied to remove the weakest connections, such that an equal network density is achieved across subjects. Recent scan–rescan studies have investigated the impact of thresholding on the reproducibility of structural brain networks in healthy subjects (Buchanan et al., 2020; Messaritaki et al., 2019; Sarwar et al., 2018; Tsai, 2018; Welton et al., 2015). Their results suggest that applying thresholds to remove false positives can improve network similarity between scan and rescan, not only in terms of graph metrics but also by replicating the same network architecture (e.g., the same set of edges and edge-weights).

Previous scan–rescan studies that tested reproducibility typically focused on datasets of healthy subjects, with state-of-the art MRI sequences, high imaging quality, and short rescan intervals (Van Essen et al., 2013). Thresholding methods may also be of value to increase consistency in network architecture in datasets of patient populations. In this setting, diffusion-based network studies are primarily used to study disease effects, cross-sectionally and over time. Before application in such clinical studies, it is essential to understand if thresholding indeed also produces more consistent network architecture in scans from patients, acquired in a clinical setting, containing various degrees of pathology, and across longer rescan intervals where further pathology has likely occurred. It is also important to determine if gain in network architecture consistency in

this setting does not come at the cost of reduced sensitivity to detect disease effects, reflected in interindividual variation in diffusion metrics, and disease-related network changes over time.

In this study, we therefore investigated whether thresholding methods that were previously shown to improve reproducibility in repeated scans of healthy young subjects also generate more consistent network architectures (e.g., the same set edges, edge-weighs, and hubs-scores) in patients who were scanned over longer time periods. To this end, we used longitudinal data of patients with cerebral small vessel disease (SVD), a condition known to affect cerebral white matter integrity that is often investigated with network analysis (Lawrence et al., 2018; Reijmer et al., 2013). In addition, we evaluated in these patients how thresholding affects sensitivity to disease effects, which was defined as (1) interindividual variation in network measures often examined in SVD (e.g., global efficiency and node strength) and (2) differences in global efficient and node strength between patients with low versus high SVD disease burden. We focused on two thresholding methods commonly applied in brain network studies: absolute thresholding and fixed-density thresholding, both of which remove low-weight connections and allow analysis of scan–rescan reproducibility on an individual patient level. As a frame of reference, we processed a dataset of healthy controls with the same methodology.

2 MATERIALS AND METHODS

2.1 Dataset 1 - Memory clinic patients with SVD

We included 228 patients from the Parelsnoer longitudinal study (Aalten et al., 2014). Patients who were referred to the memory-clinic of the UMC Utrecht for evaluation of cognitive problems, with a clinical dementia rating scale (CDR) (Morris, 1993) score of 0, .5, or 1, and a Mini Mental State Examination (MMSE) (Folstein et al., 1975) of 20 or higher were eligible. Exclusion criteria were: normal pressure hydrocephalus, Morbus Huntington, recent transient ischemic attack (TIA) or cerebrovascular accident (CVA) (<2 years), TIA/CVA followed by cognitive decline (within 3 months), history of major psychiatric disease or brain disease other than neurodegeneration or vascular disease, causing cognitive decline (e.g., brain tumour, epilepsy). Patients were eligible for the current analysis if they had a structural MRI and DTI scan at baseline and after ≈2-year follow-up visit date (N = 90). We additionally excluded three patients with cognitive complaints due to a diagnosis other than SVD or Alzheimer's disease (AD) to obtain a more homogenous study sample and one patient who was an extreme outlier in the network analysis. Thus, the total number of subjects included in the analysis was 86 (59% male). The follow-up time ranged between 22 and 35 months (mean ± SD: 27 ± 3 months) and age of the patients varied between 56 and 86 years (mean ± SD: 73 ± 7 years).

MRI data were acquired on a 3 tesla Philips scanner (Achieva, Philips, Best, the Netherlands) using a standardized clinical protocol that included a 3D T1-weighted image and a diffusion

weighted sequence. T1-weighted scans were acquired with a voxel size of 1 mm^3. DWI scans had an isotropic acquisition voxel size of 2.50 mm^3, 45 diffusion-sensitizing gradients with a b-value of 1200 s/mm2, and 1 b = 0 s/mm^2. Fluid-attenuated inversion recovery (FLAIR; TR/TE/inversion time: 11,000/125/2800 ms) were also obtained. The study was approved by the institutional review board of the UMC Utrecht, and all participants provided written informed consent prior to any research procedure.

2.2 Dataset 2 – Reference data of healthy young adults

As a frame of reference, we included a second dataset with repeated scans from healthy young adults from the Human Connectome Project (HCP, Van Essen et al., 2013). Previous studies have already tested the effects of thresholding using on this dataset but since network reconstruction pipelines always differ slightly across studies. We included these controls in our study to have a high-quality reference, reconstructed with the exact same software packages, and tractography algorithm as our patient data. We selected 44 healthy participants (32% male) with scan–rescan DWI and T1-weighted images. The rescan interval ranged between 1.5 and 11 months (mean ± SD: 4.7 ± 2 months) and the age of the participants varied between 22 and 35 years old. MRI was acquired on a Siemens Skyra 3 tesla scanner (Siemens, Erlangen, Germany). T1-weighted images had an isotropic voxel size of 1.25 mm^3. The multi-shell DWI were acquired with an isotropic voxel size of 1.25 mm^3 and three diffusion weightings (b-values: 1000, 2000, and 3000 s/mm^2). For each b-value, 90 diffusion-sensitizing gradients directions were measured. Additionally, 18 images with no diffusion weighting (b-values = 0 s/mm2) were obtained. Here, we selected only a single shell (b-value 1000 s/mm2), since it was more comparable to the patient dataset described above.

2.3 Diffusion processing and fibre tractography

All DWI scans were processed using ExploreDTI version 4.8.6 (Leemans et al., 2009) running on MATLAB R2018a (MATLAB and Statistics Toolbox Release 2018a, The MathWorks, Inc., Natick, MA, USA). Data were corrected for signal drift (Vos et al., 2017), eddy currents, subject motion with rotation of the B-matrix (Leemans & Jones, 2009), and susceptibility distortions (Veraart et al., 2013). The DWI volumes were nonlinearly registered to the T1 images prior to estimation of the diffusion tensors. Diffusion tensors were estimated using a robust method to account for outliers (Tax et al., 2015), and fibre tracts were reconstructed using deterministic fibre tractography. Seed points were distributed uniformly throughout the whole brain with 2 mm isotropic resolution. The streamlines were propagated using integration over fibre orientation distributions (FOD), with a step size of 1 mm. The orientation distributions were inferred using constrained spherical deconvolution (CSD) with a maximum harmonic order (l-max) of 6 (Jeurissen et al., 2011). Fibre tracking was terminated when streamlines entered a voxel with FOD < .1, or when the deflection angle between two successive 1 mm steps was > 45°. When tractography was concluded, streamlines with a length outside of the range between 10 and 500 mm were excluded.

2.4 Network definition

Figure 1 illustrates the processing steps for network definition. The T1-weighted scans were pre-processed using the FMRIB Software Library (FSL v5.0, Smith et al., 2004) and the Computational Anatomy Toolbox (CAT12, http://dbm.neuro.uni-jena.de/cat). First, each subject's T1 image was skull-stripped using FSL BET (Smith et al., 2004). Next, the gray matter volume was parcellated into 90 cortical and subcortical regions of interest (ROIs) using the automated anatomical labelling (AAL) atlas (Tzourio-Mazoyer et al., 2002). The parcellations were performed in the native T1 space, with the AAL template being nonlinearly registered to each subject's T1 image. The parcellated regions and the tractography data were combined to reconstruct the whole-brain network. Each ROI represented a node in the network, and two nodes were considered to be connected when they contained the endpoints of at least one streamline, resulting in a 90 × 90 binary connectivity matrices. We also computed three weighted three matrices, where the edges were weighted by the number of streamlines (NOS) connecting the two nodes, the mean diffusivity (MD), and the fractional anisotropy (FA).

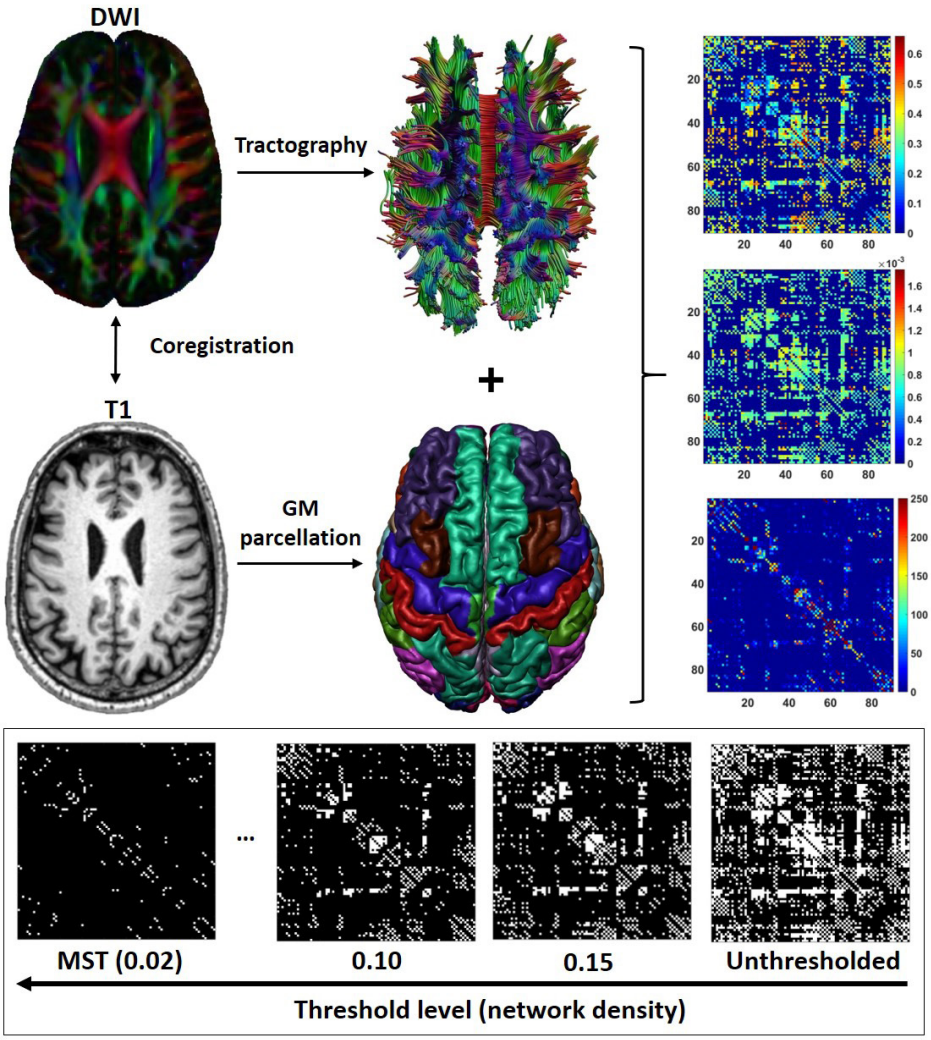


Figure 1: Network definition and thresholding. First, the DWI images were reregistered to the T1 and corrected for subject motion and artefacts. This was followed by fibre tractography and gray matter (GM) parcellation. The tractography image and segmented brain regions were combined to obtain 90×90 connectivity matrices weighted by fractional anisotropy (FA), mean diffusivity (MD), and number of streamlines (NOS). The NOS matrix was used for thresholding. The thresholded networks obtained at each threshold level were subsequently weighted by FA and MD and used in further analysis.

2.5 Network thresholding

Thresholding is frequently applied after network reconstruction, aiming to reduce the number of false-positive connections. In this work, we employed the two most common weight-based thresholding strategies: fixed-density thresholding and absolute thresholding. The fixed-density approach involved removing the edges with the lowest NOS until an equal density was achieved for all subjects. Network density is defined as the proportion of actual connections in the network, relative to all possible connections. For example, when a network has a density of 0.15, it means that 15% of all possible connections were detected in that network. To ensure that the networks did not become disconnected after thresholding we incorporated the minimal spanning tree (MST), an acyclic subgraph that connects all N nodes in the network (Tewarie et al., 2015). The MST is computed at the beginning of the thresholding step by selecting only edges with the highest NOS unless an edge forms a cycle. When all nodes are connected, the MST has N – 1 connections and a density of 2/N (≈ 0.02), with N being the number of nodes (N = 90 nodes in our case).

Using the MST as a starting point, fixed-density thresholding is applied by adding more edges to the network (from strongest to weakest weights) until a certain density is achieved. We varied the density level between the density of the MST (0.02) and the mean density of all unthresholded networks (density = 0.40) in steps of .01. Note that the stronger the threshold level, the lower the network density. The absolute thresholding approach involved removing all edges with a weight below an absolute number of streamlines. We varied the absolute threshold level between 1 and 40 streamlines in steps of 1. In this case, the stronger the threshold, the higher the number of streamlines removed. For example, a threshold of 20 streamlines means that that have fewer than 20 streamlines are removed from the network. This thresholding method does not ensure that nodes will not become disconnected or that networks of different subjects will have the same density after thresholding.

2.6 WMH volume segmentation

WMH hyperintensity volumes were segmented from the FLAIR images using and automated pipeline, kNN-TTP (Steenwijk et al., 2013).

2.7 Consistency of network architecture

To examine the consistency of network architecture between scan and rescan, we focused on characteristics that represent the building blocks of structural networks, such as edges detected, edge-weight distribution, and hub-scores (node degree and betweenness centrality).

2.7.1 Similarity in edges detected

The most direct manner of measuring similarity between networks of scan and rescan is to overlap the edges detected at both time points. Using the binary connectivity matrix, we computed the dice similarity coefficient between edges detected at scan and rescan:

$$dice = \frac{2\,|scan \cap rescan|}{|scan| + |rescan|} \qquad (1)$$

Here, the numerator represent edges in common between the two scans, whereas |scan| and |rescan| represent unique edges of scan and rescan, respectively. The dice coefficient ranges from 0 to 1, with 0 indicating no overlap and 1 representing a complete overlap between the two sets of edges.

2.7.2 Similarity in edge-weight

We also evaluated whether the edges detected in both scans have similar weights, namely the number of streamlines (i.e., the weight used to determine which edges should be retained or removed). Thus, we first computed edges in common between scan and rescan and calculated the agreement in edge-weight using the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979). The ICC was originally created to assess the reliability of multiple ratters measuring the same variable, but it is also often utilized in network studies to assess the consistency of graph measure over multiple sessions (Andreotti et al., 2014; Buchanan et al., 2014; Messaritaki et al., 2019):

$$\mathrm{ICC} = \frac{\mathrm{MS}_b - \mathrm{MS}_w}{\mathrm{MS}_b + (k-1)\mathrm{MS}_w} \qquad (2)$$

where $\mathrm{MS}_{\mathbf{b}}$ is the between-subject variance, MS_w represents within-subject variance, and k is the number of repeated measurements. ICC values range between 0 and 1 and are typically interpreted as poor (<0.40), fair (0.40–0.59), good (0.60–0.74), and excellent (>0.75) (Cicchetti, 1994; Wang et al., 2019; Yuan et al., 2019).

2.7.3 Similarity in hub score

Another relevant feature of network architecture is the location of hub nodes. Hubs are topologically central regions and are positioned to make strong contributions for global network function (van den Heuvel & Sporns, 2013). The two most common graph metrics used to define node importance and to identify hubs are node degree and node betweenness centrality. Degree refers to the number of connections that link one node to adjacent nodes. Betweenness centrality is defined as the fraction of all shortest paths in the network that pass through a given node (Bullmore & Sporns, 2009). These nodes with high "hub-score" (i.e., high degree and/or betweenness centrality) contribute to an efficient communication between distant brain regions. We compared the similarity in hub scores of all nodes between scan and rescan by computing the ICC of betweenness centrality and ICC of degree.

2.8 Interindividual variation and sensitivity to changes over time in patients

Here, we examined how thresholding affects the natural interindividual variation in the data, necessary to test associations with external variables and perform group comparisons (Bagarinao et al., 2019). We focused on metrics such as global efficiency and FA- and MD-weighted node strength. Global efficiency was defined as the inverse of the average shortest path length and quantifies how efficiently information is exchanged over the network (Rubinov & Sporns, 2010). Node strength was defined as the average FA or MD of all edges connected to a node. Clearly, disturbances of these network metrics are not specific to only SVD. Moreover, a range of other metrics exist.

We choose these particular measures because they are known to be affected by SVD (Lawrence et al., 2014; Reijmer et al., 2013). Moreover, these metrics (i.e., lower global efficiency and FA, and a higher MD) are known to be related to disease burden and progression over time. Patients with larger WMH volumes show lower global efficiency and FA, and a higher MD, and as the disease progresses over time, global efficiency, FA and MD are expected to decline further (Tuladhar et al., 2020). Therefore, we also tested among the patients how thresholding affected the effect size of the difference in diffusion metrics between high and low disease burden and sensitivity to detect changes over time. The paragraphs below describe how we evaluated the interindividual variation of these metrics and sensitivity to detect time and group effects for different threshold levels.

2.8.1 Assessing interindividual variation

For each threshold level, global efficiency and node strength were calculated and transformed to z-scores. Follow-up z-scores were calculated using the mean and standard deviation of baseline. To assess whether thresholding affects interindividual variation, we computed a correlation matrix containing Pearson's correlation coefficients between z-scores of different threshold levels. Figure 2 shows an example of how this correlation matrix was calculated for global efficiency (GE). First, global efficiency z-scores were calculated for each subject (S) and for each threshold level (T). Then, z-scores of one threshold level were correlated with z-scores of another threshold level, yielding a correlation coefficient (r). The strength of this correlation indicates whether the interindividual variation in global efficiency is preserved between threshold levels. In other words, if baseline scores change between threshold levels, this results in low correlation coefficients. We calculated r between all combinations thresholds to fill the correlation matrix. In this manner, it is possible to observe which threshold levels generate substantially different global efficiency values.

We also computed the same correlation matrix for rate of the change in global efficiency over time (i.e., baseline minus follow-up), to understand whether thresholding affects the estimated individual rate of change over time. Essentially, for a hypothetical subject with a baseline global efficiency of 1.5×standard deviation (SD) relative to the group mean, if the interindividual

variation is not affected, the same score should be obtained for that subject when a different threshold is used. Similarly, if the global efficiency of that subject declined –0.5 SD from baseline to follow-up, the same rate of decline should be observed at different thresholds. The range of thresholds that shows the highest r values thus represent the thresholds were sensitivity to interindividual variation is optimal. Differences in follow-up time between subjects were not adjusted in the analyses, because such differences only further contribute to interindividual variation, and our intention was to assess that variation.

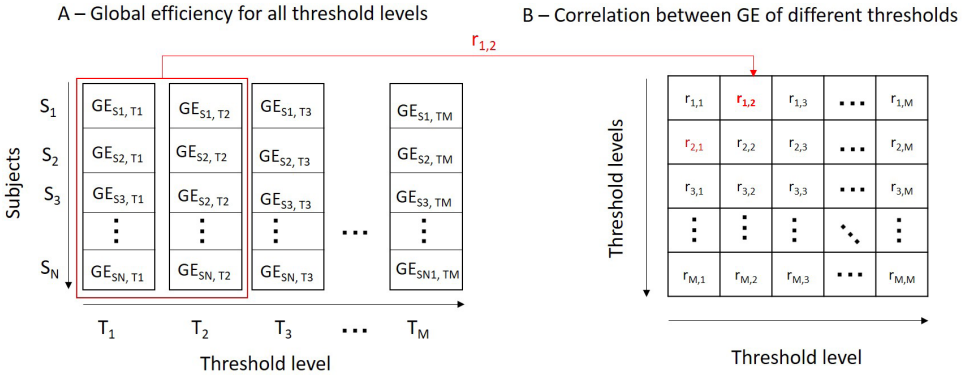


Figure 2: Correlation matrix of global efficiency values of different thresholds. A: First, for each threshold level (T), z-scores of global efficiencies (GE) were calculated for all subjects (S). Then, to examine whether the inter-individual variation changes between threshold levels (e.g., level 1 and 2), we calculated the Pearson correlation coefficient between global efficiency of those threshold levels ($r_{1,2}$), resulting correlation matrix containing correlations between all pairs of threshold levels. B: Correlation coefficients were calculated for all combination of thresholds, resulting in a correlation matrix. Note that this matrix is symmetric since $r_{1,2} = r_{2,1}$.

2.8.2 Assessing changes over time

To examine whether thresholding affects the sensitivity to detect network changes over time in patients, we used mixed ANOVA to compare baseline versus follow-up global efficiency and node strength. Patients were stratified into two groups using a median split of WMH volume. For this analysis, 13 patients were excluded due to lack of FLAIR images for the segmentation of WMH, resulting in 73 patients. We evaluated whether the sensitivity to detect an effect of time (within-subject factor), effect of group (between-subject factor), and interaction time × group is preserved across thresholds.

RESULTS

3.1 Consistency of network architecture

Figure 3 shows the effect of thresholding on the consistency of network architecture in patients. In each panel, the first plot contains similarity scores for different characteristics of the network architecture, and the second plot shows the number of nodes that remain in the network at each threshold level. Quantitate values are shown in Table S1. Results with the reference dataset of controls are shown in the Supporting Information (Figure S1 and Table S2).

3.1.1 Similarity in edges detected

In the patients, the dice similarity of edges was 0.70 before thresholding. When fixed density thresholding was applied, this score increased with stronger thresholds (i.e., with lower densities), reaching a maximum of 0.76 (at a density = 0.08, $p < 0.001$ compared to unthresholded, Figure 3A, blue line, Table S1). For densities lower than 0.08, dice scores decreased sharply to .67 (density = 0.02). As expected, the number of nodes (N = 90) in the network did not change with fixed-density thresholding. Regarding absolute thresholding, the dice score also increased with stronger thresholds (i.e., with larger number of streamlines removed), from 0.70 (unthresholded) to a maximum of 0.76 (# streamlines removed = 40, $p < 0.001$, Figure 3B, blue, Table S1). With absolute thresholding, the maximum dice score was produced by the strongest threshold. However, at this threshold level, only 38 of the initial 90 nodes remained in the network. Overall, these results indicate that compared to not applying any threshold whatsoever, thresholding generates a more similar set of edges between scan and rescan.

In controls, thresholding had a similar effect on dice similarity, albeit with higher scores than in patients as expected, because of higher quality scans and absence of pathology (Figure S1A and B, blue, Table S2).

3.1.2 Similarity in edge-weight

The ICC of edge-weight in patients was .71 for unthresholded networks (Figure 3A, black). After fixed-density thresholding, ICC scores increased with stronger thresholds to a maximum of 0.75 (density = 0.12, $p < 0.001$, Figure 3A, black). When absolute thresholding was applied, the ICC also increased with stronger thresholds, reaching a maximum of 0.76, again at the strongest threshold level (# streamlines removed = 40 streamlines, Figure 3B, black, Table S1). These results indicate that the edges retained after thresholding have more consistent weight distributions between scan and rescan. Similar results were observed in controls, but with higher ICC scores (Figure S1).

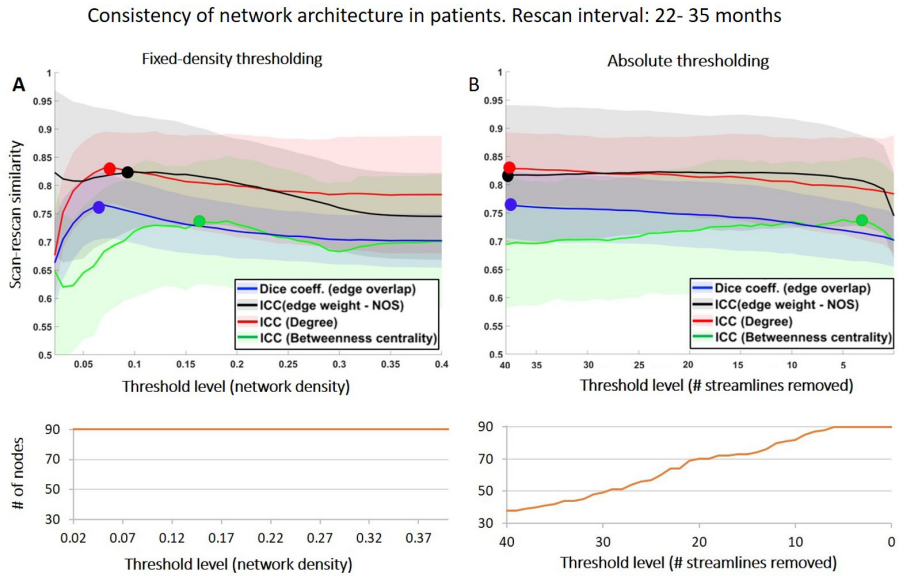


Figure 3: Consistency of network architecture between scan and rescan in patients (rescan interval = 22 - 35 months). In each plot, the x-axis represents the threshold level, with the strength of thresholding increasing from right to left. Note that for the fixed-density approach the stronger the threshold the lower the density, whereas for the absolute approach, the stronger the threshold, the higher the number of streamlines removed. The curve in orange under each plot shows the number of nodes that remain in the network after thresholding. The initial number of nodes for the unthresholded network was N = 90 nodes. Blue: dice similarity of edges, black: ICC of edge-weights; red: ICC of degree; green: ICC of betweenness centrality. The markers highlight maximum value in each curve, and the shaded areas represent the standard deviation.

3.1.3 Similarity in hub score

The ICC degree also increased after thresholding, from 0.78 (unthresholded) to 0.83 when fixed-density thresholding was used (density = 0.08, $p < 0.001$, Figure 3A, red, Table S1), whereas the ICC of betweenness centrality showed an unstable profile and did not increase significantly after thresholding (Figure 3a, green). Regarding absolute thresholding, the ICC degree also increased from 0.78 to a maximum of 0.83 (Figure 3B, red), and ICC-betweenness centrality did not increase with stronger thresholds. For controls, the effect of thresholding on hubs scores was analogous (see Figure S1), albeit with higher scores.

3.2 Interindividual variation and sensitivity to changes over time in patients

The analyses assessing interindividual variation and sensitivity to changes in global efficiency, MD- and FA-weighted node strength in patients are shown in Figures 4-6. In each figure, "panel a" contains spaghetti plots summarizing baseline and follow-up z-scores, as well as correlations between baseline and follow-up, within each threshold level. From a biological perspective, these correlations indicate whether patients maintain their relative position in group from baseline to follow-up, which would then be reflected in parallel lines in the spaghetti plots, and strong

correlations. Panel B (left) shows correlation matrices between baseline values of different threshold levels. These correlations indicate whether individual patients maintain their relative baseline scores across different thresholds. Threshold levels that do not disturb this interindividual variation of baseline scores should show high correlations. Likewise, panel B (right) shows a correlation matrix but for the rate of decline over time across thresholds, indicating whether individual patients maintain the same rate of decline between thresholds. Panel C depicts the sensitivity to detect time and group effects in patients stratified by WMH volume for different threshold levels. If a statistically significant change over time is found, this will be reflected in an F-value (time) > critical value. The magnitude of this effect is given by the Cohen's d. Similarly, if there is significant difference between patients with low and high WMH volume, this group-effect is given by the F-value (group) and corresponding effect size.

Since absolute thresholding produced a large number of disconnected nodes, some brain regions got excluded from the analysis, not allowing us to always evaluate global efficiency over the same network of 90 nodes. It is well known that global network measures are highly dependent on the number of nodes (van Wijk et al., 2010). Therefore, for the analysis regarding the sensitivity to disease effects, we focused on the fixed-density approach where the size of the network is maintained regardless of the threshold level, whereas the results for absolute thresholding are shown in the Supporting Information (Figures S1–S3).

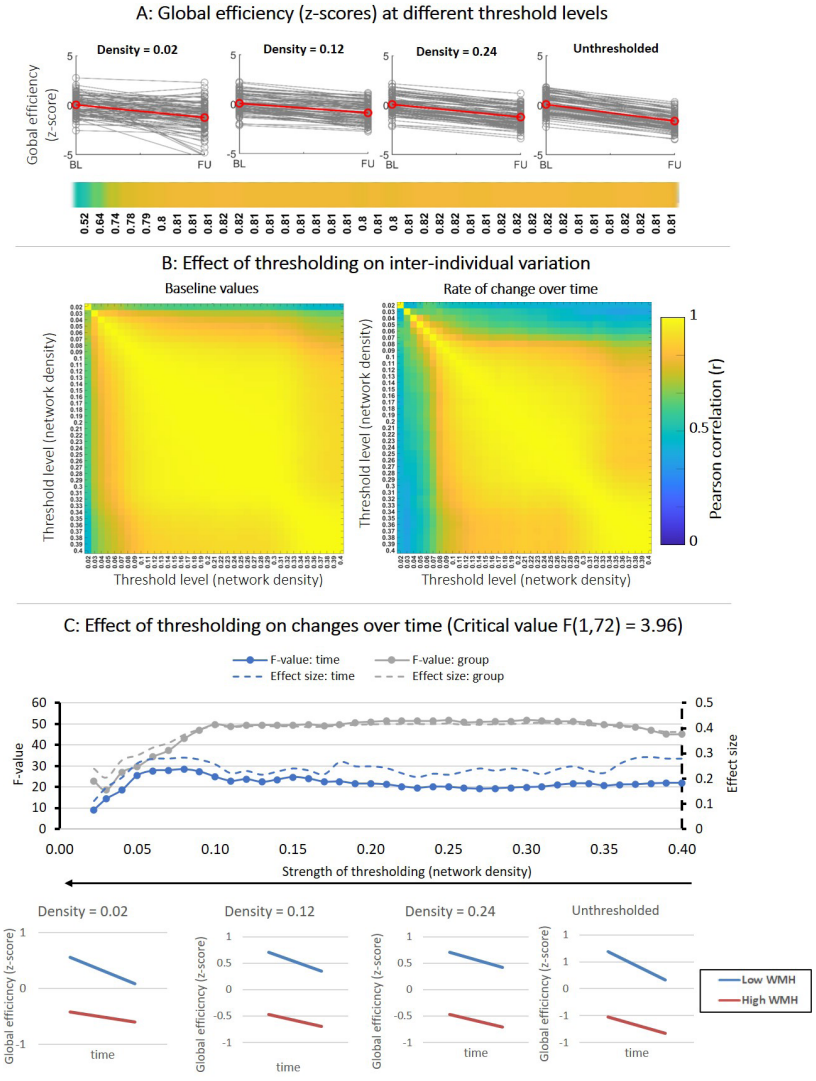


Figure 4: Effect of fixed-density thresholding on global efficiency. A: Spaghetti plots showing baseline and follow-up z-scores for different threshold levels. The red line represents the group average. The horizontal colour bar shows correlations between baseline and follow-up scores for each threshold level. B: (Left) Correlation matrix containing Pearson correlations between baseline global efficiency scores of different threshold levels. High correlations indicate that the baseline scores are similar between threshold levels. (Right) Correlations between the rate of change over time obtained at different threshold levels. High correlations indicate that the individual rate of change is similar between threshold levels. C: Impact of thresholding on the sensitivity to detect changes over time and group differences in patients stratified by WMH volume. (Top) F-values and effect sizes were calculated using mixed ANOVA with time as within subject factor and group (low vs. high WMH) as between-subject factor. Left axis corresponds to F-values and right axis represents the effect sizes for each effect: time (blue), group (gray). (Bottom) Average change over time in global efficiency for patients with low vs. high WMH volume. Patients with high WMH have lower efficiency scores and both groups declined over time.

3.2.1 Global efficiency

Figure 4 shows results for global efficiency. In Figure 4A, the spaghetti plots indicate that before thresholding there is a clear variation in global efficiency between subjects, with an apparent decline from baseline to follow-up. The lines of the subject initially run in parallel before thresholding, reflected in high a correlation between baseline and follow-up scores ($r = 0.81$). After thresholding, this variation of baseline and follow-up scores was maintained ($r \approx 0.8$). The relation between baseline and follow-up scores was only disturbed when networks were thresholded to densities < 0.05, implying that interindividual differences between subjects is no longer maintained.

In Figure 4b, we quantified whether the interindividual variation of subjects and the individual rate of change over time are maintained between different threshold levels. Regarding baseline values (left matrix), z-scores of global efficiency remained robust between density thresholds of 0.40 and 0.05 ($r > 0.9$). In other words, the distance of each subject to the group mean at baseline was unchanged, as long as networks were not thresholded to densities below 0.05. The same was true for the change over time (right matrix). The individual rate of change from baseline to follow-up was unaffected for densities between 0.40 and 0.10 ($r > 0.9$). Overall, these correlations analyses show that global efficiency scores of individual subjects do not change after thresholding ($0.10 < \text{density} < 0.40$).

In Figure 4C, we statistically tested whether the apparent decline over time in global efficacy observed in the spaghetti plots was statistically significant, and whether patients with low WMH differed from high WMH volume. Before thresholding (density = 0.40), there was significant effect of time (blue curve), indicating a decline in global efficiency from baseline to follow-up ($F(1, 72) = 22$; $p < 0.01$; Cohen's $d = 0.29$). This decline is in line with previous reports (Tuladhar et al., 2020). Regarding the group effect (gray curve), patients with larger WMH volumes had a significantly lower global efficiency ($F(1, 72) = 45$; $p < 0.01$; Cohen's $d = 0.4$). There was no interaction of time × group, with both groups of patients declining at the same rate. After thresholding, the sensitivity to detect time and group effects was robust, with F-values and effect sizes remaining relatively constant for a wide range of densities between 0.40 and 0.10. This result indicates that these threshold levels preserve the sensitivity to effects of time and group observed originally before thresholding. The absolute thresholding method resulted in larger changes on the interindividual variation and sensitivity do detect time and group effects (Figure S1). Only a narrow range of thresholds preserved the sensitivity to detect time and group-effects.

3.2.2 MD-weighted node strength

Figure 5 describes the same analysis as Figure 4, but for MD-weighted node strength. By observing the spaghetti plots on Figure 5a, MD appears unchanged from baseline to follow-up

on a group level.

However, the correlation matrices on Figure 5B reveal that between a density of 0.40 and 0.30, MD scores remain initially unchanged ($r > 0.9$, note the block of high correlations on the bottom right of the matrix). When stronger thresholds were applied (i.e., when more noisy connections were removed, densities < 0.30), MD scores obtained at these threshold levels no longer resemble those obtained for thresholds between .40 and 0.30 but did not change further for the remaining thresholds ($r > 0.9$, for densities between 0.30 and 0.02).

In Figure 5C, we tested the sensitivity to detect changes over time. Before thresholding, there was a significant but small effect of time on MD ($F(1, 72) = 12$; $p < 0.01$; Cohen's $d = 0.1$). The group-effect and the interaction time × group were not significant. After thresholding, the sensitivity to detect an effect of time was increased (i.e., larger F-values) and highest for density thresholds around 0.12. Note that these were the threshold levels that also improved the consistency of network architecture over time. This indicates that removing noisy connections improves the sensitivity to detect small changes within individuals over time in local weight-based metrics such as MD. Absolute thresholding produced larger changes on interindividual variation but did not eliminate the sensitivity to effects of time and group (Figure S2).

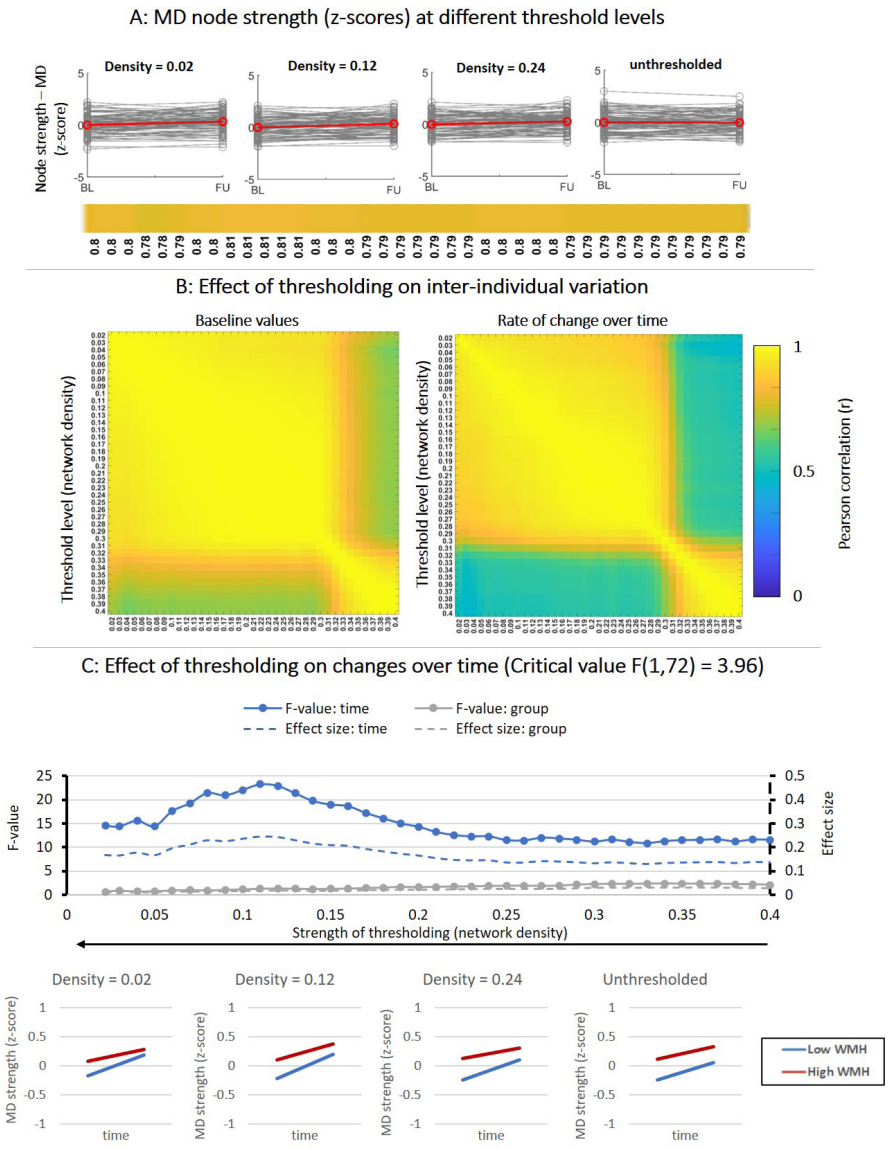


Figure 5: Effect of fixed-density thresholding on MD-weighted node strength. A: Baseline and follow-up z-scores for different threshold levels. The horizontal colour bar shows correlations between baseline and follow-up scores within each threshold level. B: (Left) Correlation matrix containing Pearson correlations coefficients between baseline scores of different threshold levels. (Right) Correlations between rate of change over time obtained at different threshold levels. C: Impact of thresholding on the sensitivity to detect changes over time and group-differences in patients stratified by WMH volume. (Top) F-values and effect sizes for the effects of time (blue), group (gray). (Bottom) Average change over time in MD-weighted strength for patients with low vs. high WMH volume.

3.2.3 FA-weighted node strength

Figure 6 shows the results for FA-weighted node strength. Thresholding appears to have a small effect on the relation between baseline and follow-up FA values as illustrated by the spaghetti plots and the correlations between baseline and follow-up (Figure 6a).

The correlations matrices (Figure 6B) further show that the interindividual variation of baseline values (left matrix) remained robust for all threshold levels between 0.37 and 0.02 ($r > 0.9$). In other words, the position of each individual subject relative to the group mean did not change for thresholds between 0.37 and 0.02. The same result was obtained for the individual rate of change over time (right matrix).

Figure 6C shows that before thresholding no significant time and group effects were detected. Thresholding did not improve the sensitivity and effect of time but did improve the sensitivity do detect a small effect of group for densities < 0.30: ($F(1, 72) = 5$; $p < 0.05$; Cohen's $d = 0.05$).

Again, absolute thresholding caused larger changes on interindividual variation but also improved the detection group-effects after the first threshold levels (Figure S2).

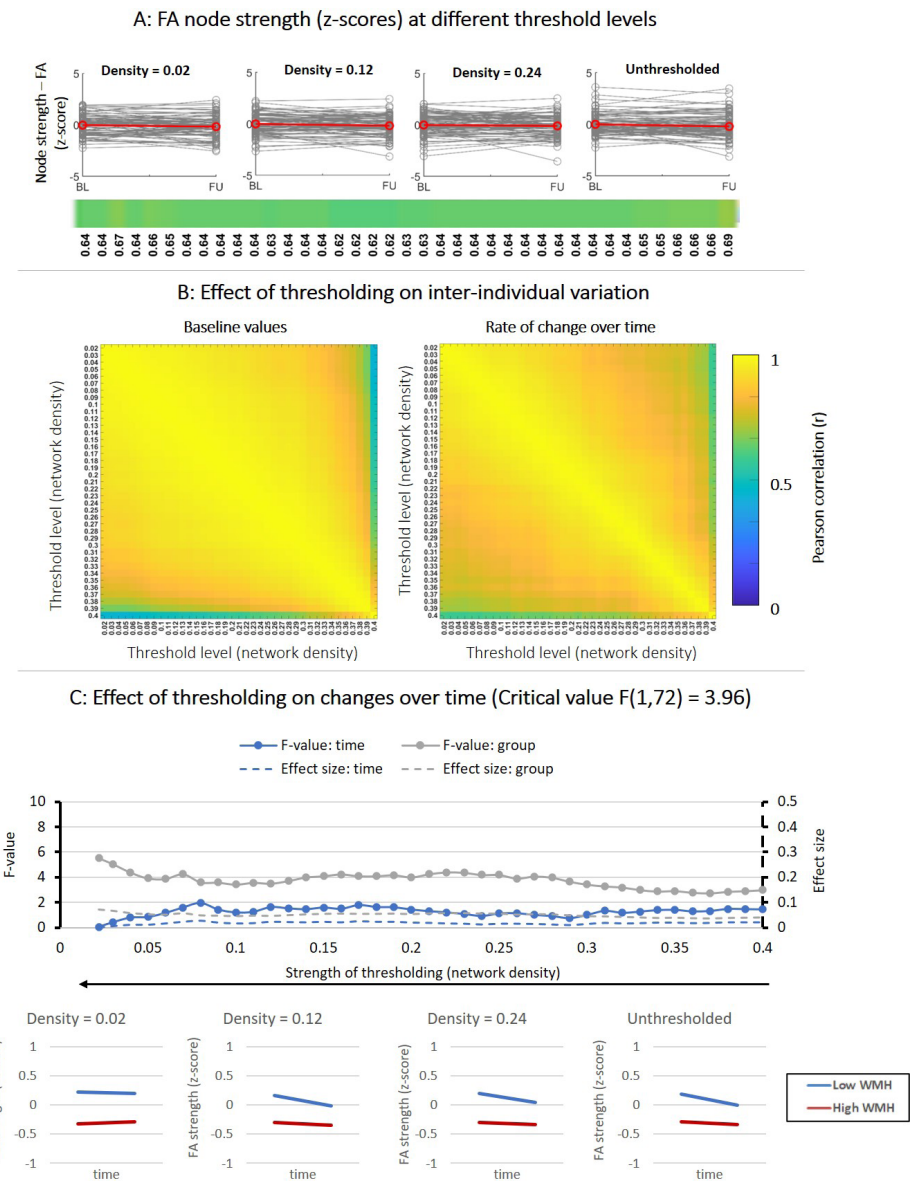


Figure 6: Effect of fixed-density thresholding on FA-weighted node strength. A: Baseline and follow-up z-scores for different threshold levels. The horizontal colour bar shows correlations between baseline and follow-up scores within each threshold level. B: (Left) Correlation matrix containing Pearson correlations coefficients between baseline scores of different threshold levels. (Right) Correlations between rate of change over time obtained at different threshold levels. C: Impact of thresholding on the sensitivity to detect changes over time and group-differences in patients stratified by WMH volume. (Top) F-values and effect sizes for the effect s of time (blue), group (gray), interaction term (orange). (Bottom) Average change over time in FA-weighted strength for patients with low vs. high WMH volume.

4 DISCUSSION

In this work, we evaluated the impact of thresholding on scan–rescan brain networks of patients with SVD to assess how thresholds that improve scan–rescan network reproducibility in healthy young subjects affect (1) consistency in network architecture in these patients over a longer time period and (2) sensitivity to detect biological effects. Our results indicate that threshold levels that improve the reproducibility in controls also generate more consistent network architecture over time in patients. The similarity between scan and rescan for characteristics such as the location of edges detected, edge-weights, and hub scores improved after thresholding. We also showed that the natural interindividual variation in outcome measures used to assess disease effect is preserved within threshold levels where the network architecture is consistent. Furthermore, the sensitivity to detect statistical group differences between patients with low vs. high WMH burden was maintained.

Preceding our work, several studies had examined the effect of thresholding on reproducibility of network architecture, including the binary topology, edges detected, edge-weights, and graph metrics in healthy, mostly young, controls (Andreotti et al., 2014; Buchanan et al., 2014; Owen et al., 2013). As expected, our results with controls were in line with studies that explicitly showed that networks become more reproducible after thresholding (e.g., Buchanan et al., 2020; Messaritaki et al., 2019; Roine et al., 2019). In patients where disease effects are monitored over time, the rescan interval is typically much longer. Therefore, the same thresholding methods and threshold levels that are reported to improve reproducibility in controls might not directly apply. We used SVD as an exemplar condition, because network metrics have been shown to be relevant for this disease, but also because SVD-related brain injury such as white matter hyperintensities, brain atrophy, and enlarged ventricles can impact the performance of tractography and network reconstruction. Current fibre tractography methods do not explicitly account for such factors, which can lead to erroneous estimations of white matter pathways and presence of more false positives in the network. Our results with patients show that the consistency of edges detected and their respective weights are improved when both thresholding methods are used, suggesting that the edges removed are noisy connections with a more random weight distribution (Messaritaki et al., 2019; Zalesky et al., 2016). Thresholding also improved the ICC of degree, meaning that if we were to define hubs nodes based on degree, a more consistent set of hubs would be detected between scans. Notably, the betweenness centrality was less consistent between scans, which could be explained by the fact that this metric depends not only on edges directly connected to a specific node but also on edges connected to distant nodes. Since the betweenness centrality quantifies the proportion shortest paths that go through given node, removing only one edge (which can be directly or indirectly connected to that node) can have a large impact on that shortest path. Thus, the betweenness centrality is more susceptible to disruptions when edges are removed (Drakesmith et al., 2015; Segarra & Ribeiro, 2014).

The two thresholding methods had distinct threshold levels to achieve optimal reproducibility. For the fixed-density approach, similarity scores improved with decreasing density (i.e., stronger

thresholds), before drastically decreasing for densities < .05. This reproducibility profile could be explained by the proportion of false positive at low densities and by the MST (Zalesky et al., 2016). The MST was incorporated to ensure that networks remain connected and avoid fragmented nodes. Since, by definition, the MST cannot contain connections that form cycles, a certain proportion of low-weight connections (i.e., potential false positives) must be included to keep the network connected, meaning that at very low densities, the effect of these false positives is stronger. A potential disadvantage of thresholding networks to fixed-density is that it can lead to confounding effects when comparing groups or datasets with different distributions of edge weights. In a group with higher edge weights, this would lead to ignoring potentially important edges with strong weights, while in a group with lower edge weights, this would lead to including weak or potentially spurious edges (van Wijk et al., 2010). Regarding the absolute thresholding approach, the similarity scores increased with mild thresholds (2–5 streamlines). Since this approach works by removing all connections with a weight below a certain number of streamlines, only the highest weighted connections survive when strong thresholds are used. A major downside of this method, also evident in our dataset, is that it quickly creates fragmented nodes, which means that some brain regions are no longer part of the network. Furthermore, differences in brain size or absolute number of streamlines computed for each subject results in largely different networks between subjects or over time. Thus, a "one size fits all" type of threshold is not ideal. For datasets of patients with similar characteristics as those included in this study, we therefore recommend using fixed-density thresholds between 0.08 and 0.20 to achieve optimal network consistency while keeping all network nodes connected. This is also in line with previous research that estimated the density of structural connectomes to lie between 0.05 and 0.30 (Hagmann et al., 2008; Roberts et al., 2017).

One of the main arguments against thresholding networks of patients in the attempt to improve network consistency over repeated scans is that this procedure could remove biological or disease-related effects (Drakesmith et al., 2015; McColgan et al., 2018). This concern can be relevant in studies trying to identify disease effects at the level of subnetwork and/or individual connections, rather than pathological changes in large-scale brain network topology (Petersen et al., 2020). In the SVD field, this type of analyses could help understanding how diffuse and/or focal damage in certain brain areas affects cognitive function. In those scenarios, applying thresholds could erroneously cut connections that are affected by pathology. On the other hand, for global network metrics, it is more desirable to assess disease effects on networks with a more consistent scan–rescan architecture, less affected by noisy connections that can confound the results. Detection of intraindividual changes over time in patients is generally one of the more challenging tasks in terms of sensitivity. Such changes may be detected at a group level, but the individual trajectories of patients tend to get overshadowed by noise if techniques are insensitive. Therefore, analysis of interindividual variation over time can be performed with more confidence if networks are voided of noisy connections. In the second part of our analysis, we show that the concerns about removal of disease effects by thresholding may not always be justified. For global efficiency, interindividual variation of subjects at baseline and their

individual rate of change over time remained robust for fixed-density thresholds > 0.10. Before thresholding, global efficiency was able to capture significant differences between patients with low and high WMH volume, which is in line with previous research where white matter lesion load was associated with lower global efficiency (Heinen et al., 2018). After fixed-density thresholding, a wide-range of densities (0.40–0.10) preserved the sensitivity to detect these disease-related effects, suggesting that changes in global metrics can be consistently detected over multiple threshold levels (de Brito Robalo et al., 2020; Drakesmith et al., 2015). Absolute thresholding had a stronger impact interindividual variation and sensitivity to time and group effects (see Supporting Information), since this thresholding approach creates disconnected nodes during, thereby changing the size of the network and disrupting global network metrics (van Wijk et al., 2010).

Regarding MD-weighted node strength, the sensitivity to detect time effects was improved after thresholding (density ≈ 0.12). Since this network metric is based on edge weights, the measurements after thresholding are obtained over a smaller and more consistent set of connections, thereby decreasing the standard error and improving the sensitivity to detect small effect sizes. This hypothesis is also supported by previous work that examined associations between edge-weights and age and showed that connections retained in the network after thresholding were significantly more associated with age than those removed (Buchanan et al., 2020). FA-weighted node strength was lower in patients with higher WMH volume but did not significantly decline over time. These effects were not affected by thresholding.

Strengths of this work include the use of two distinct datasets, with different MRI protocols, different subject groups, and rescan intervals. In this manner, it was possible to directly test whether thresholds that improve the reproducibility in high-quality scans with short rescan intervals also have the same effect on scans of patients, with longer follow-up intervals. Our study also had some limitations. We analysed the effects of thresholding only on network metrics that have shown association with disease effects in patients with SVD, such global efficiency and node strength. Thus, our results cannot be directly generalized to all network measures and all patient populations. Further investigation on disease effects reflected by other network metrics is required (e.g., local metrics or disease effects at the level of subnetworks). Our analysis was focused on the most popular weight-based thresholding methods that directly remove connections from each individual network, without the need to create a group-level network to determine a criterion to remove connections. In this manner, test–retest consistency of network architecture can be evaluated on an individual basis. Other scan–rescan studies have examined thresholding approaches that are not based on edge-weights but rather on group-level consistency and is less biased towards the length of streamlines (Buchanan et al., 2020; Roberts et al., 2017). Their findings also point that stringency of threshold can improve network consistency. However, future work should also analyse how consistency-based thresholding improves group-level consistency of network architecture (e.g., across scanners) in similar datasets and whether disease effects are preserved. Our results could also have been influenced by the choice

of parcellation scheme (Zalesky et al., 2010), tractography algorithm (Bastiani et al., 2012), weighting scheme, among other factors. Thus, these results need to be tested using different network reconstruction pipelines.

5 CONCLUSIONS

Our study demonstrates the effects of weight-based thresholding on longitudinal brain networks of patients with SVD. We showed that thresholding, particularly with fixed-density approaches, can produce more consistent network architectures in patients scanned over longer time periods, while preserving disease-related effects. Our work sheds a light on how to make informed decisions when applying thresholds in studies with a longitudinal design and how such choices can potentially influence the statistical significance of the results. A good practice for longitudinal studies that intend to apply weight-based thresholds would be to first examine which threshold levels generate the most consistent network architecture over time and then verify if those threshold levels also preserve the interindividual variation in metrics that will be used as outcome for the study (e.g., global efficiency).

ACKNOWLEDGMENTS

This work was supported by ZonMw, The Netherlands Organisation for Health Research and Development (VICI grant 91816616 to Geert Jan Biessels). Yael D. Reijmer received funding from Alzheimer Nederland and ZonMw/Deltaplan Dementie (grant #733050503) and a Young Talent Fellowship from the Brain Center Rudolf Magnus, University Medical Center Utrecht. The research of Alexander Leemans is supported by VIDI Grant 639.072.411 from the Netherlands Organization for Scientific Research (NWO). Members of the Utrecht Vascular Cognitive Impairment (VCI) Study group involved in the present study (in alphabetical order by department): University Medical Center Utrecht, The Netherlands, Department of Neurology: E. van den Berg, J. M. Biesbroek, G. J. Biessels, M. Brundel, W. H. Bouvy, L.G. Exalto, C. J. M. Frijns, O. Groeneveld, S. M. Heringa, N. Kalsbeek, L. J. Kappelle, Y. D. Reijmer, J. Verwer; Department of Radiology/Image Sciences Institute: J. de Bresser, H. J. Kuijf, A. Leemans, P. R. Luijten, M. A. Viergever, K. L. Vincken, J. J. M. Zwanenburg; Department of Geriatrics: H. L. Koek; Hospital Diakonessenhuis Zeist, The Netherlands: M. Hamaker, R. Faaij, M. Pleizier, E. Vriens.

SUPPLEMENTARY INFORMATION

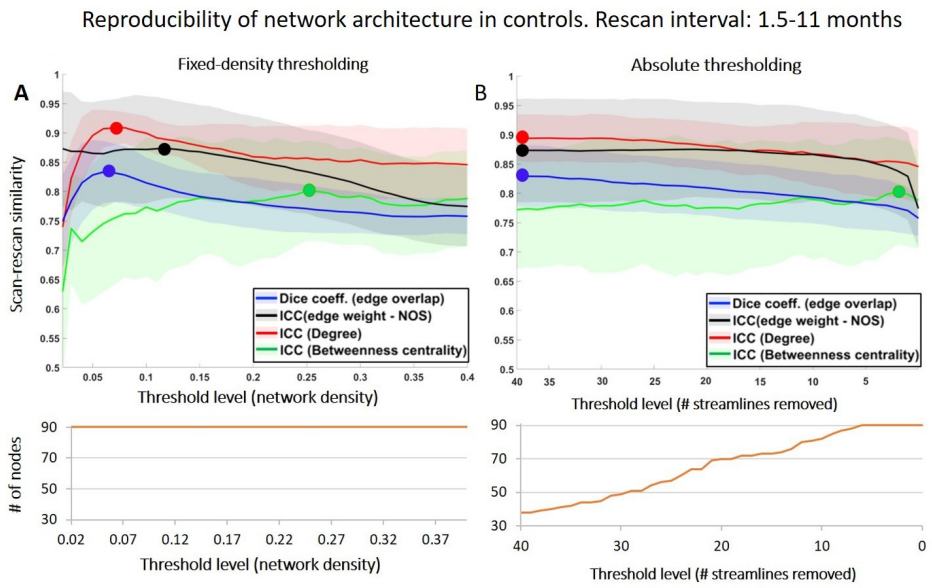


Figure S1: Reproducibility of network architecture between scan and rescan in controls (rescan interval = 1.5 – 11 months). In each plot, the x-axis represents the threshold level, with the strength of thresholding increasing from right to left. Note that for the fixed-density approach the stronger the threshold the lower the density, whereas for the absolute approach, the stronger the threshold, the higher the number of streamlines removed. The curve in orange under each plot shows the number of nodes that remain in the network after thresholding. The initial number of nodes for the unthresholded network was N = 90 nodes. Blue: dice similarity of edges, black: ICC of edge-weights; red: ICC of degree; green: ICC of betweenness centrality. The markers highlight maximum value in each curve, and the shaded areas represent the standard deviation.

Table S1: Similarity of network architecture between scan and rescan in patients, before and after thresholding.

	Metric	Unthresholded	Thresholded	Effect size (Cohen's d)	p-value
Fixed-density thresholding	Dice coefficient	0.70 ± 0.04	0.76 ± 0.06	1.5	<0.001
	ICC (edge weight – NOS)	0.71 ± 0.06	0.75 ± 0.09	0.73	<0.001
	ICC (degree)	0.78 ± 0.10	0.83 ± 0.06	0.9	<0.001
	ICC (betweenness centrality)	0.70 ± 0.12	0.70 ± 0.12	0.12	0.75
Absolute thresholding	Dice coefficient	0.70 ± 0.04	0.76 ± 0.02	1.76	<0.001
	ICC (edge weight – NOS)	0.71 ± 0.06	0.76 ± 0.08	0.75	<0.001
	ICC (degree)	0.78 ± 0.10	0.83 ± 0.13	0.7	<0.001
	ICC (betweenness centrality)	0.70 ± 0.12	0.72 ± 0.2	0.19	0.06

Effect sizes and p-values obtained from paired samples t-test.

Table S2: Similarity of network architecture between scan and rescan in controls, before and after thresholding.

	Metric	Unthresholded	Thresholded	Effect size (Cohen's d)	p-value
Fixed-density thresholding	Dice coefficient	0.76 ± 0.03	0.83 ± 0.07	2.1	<0.001
	ICC (edge weight – NOS)	0.77 ± 0.08	0.84 ± 0.10	1.5	<0.001
	ICC (degree)	0.85 ± 0.04	0.89 ± 0.07	0.81	<0.001
	ICC (betweenness centrality)	0.79 ± 0.10	0.80 ± 0.11	0.14	0.12
Absolute thresholding	Dice coefficient	0.76 ± 0.03	0.82 ± 0.05	1.79	<0.001
	ICC (edge weight – NOS)	0.77 ± 0.08	0.84 ± 0.12	1.2	<0.001
	ICC (degree)	0.85 ± 0.04	0.89 ± 0.06	0.77	<0.001
	ICC (betweenness centrality)	0.79 0.10	0.80 ± 0.16	0.1	0.34

Effect sizes and p-values obtained from paired samples t-test.

Inter-individual variation and sensitivity to changes over time – absolute thresholding

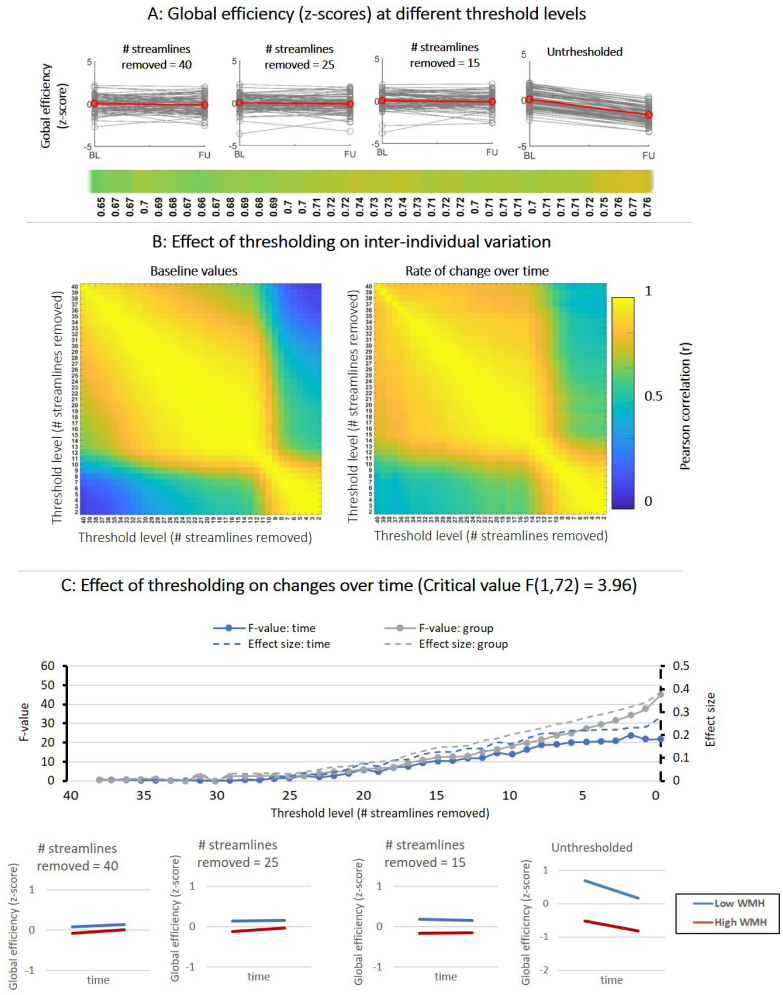


Figure S2: Effect of absolute thresholding on global efficiency. A: Spaghetti plots showing baseline and follow-up z-scores for different threshold levels. The red line represents the group average. The horizontal colour bar shows correlations between baseline and follow-up scores for each threshold level. B: (Left) Correlation matrix containing Pearson correlations between baseline global efficiency scores of different threshold levels. High correlations indicate that the baseline scores are similar between threshold levels. (Right) Correlations between the rate of change over time obtained at different threshold levels. High correlations indicate that the individual rate of change is similar between threshold levels. C: Impact of thresholding on the sensitivity to detect changes over time and group-differences in patients stratified by WMH volume. (Top) F-values and effect sizes were calculated using mixed ANOVA with time as within subject factor and group (low vs. high WMH) as between-subject factor. Left axis corresponds to F-values and right axis represents the effect sizes for each effect: time (blue), group (gray). (Bottom) Average change over time in global efficiency for patients with low vs. high WMH volume. Patients with high WMH have lower efficiency scores and both groups declined over time.

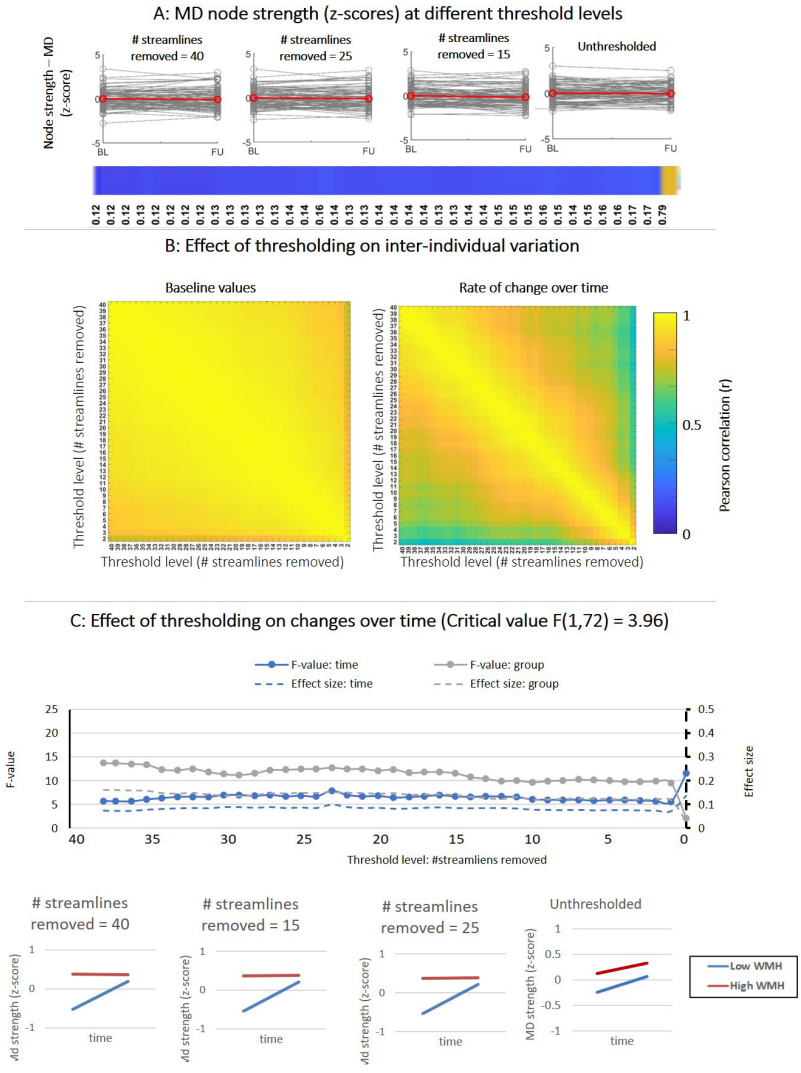


Figure S3: Effect of absolute thresholding on MD-weighted node strength. A: Baseline and follow-up z-scores for different threshold levels. B: (Left) Correlation matrix containing Pearson correlations coefficients between baseline scores of different threshold levels. (Right) Correlations between rate of change over time obtained at different threshold levels. C: Impact of thresholding on the sensitivity to detect changes over time and group-differences in patients stratified by WMH volume. (Top) F-values and effect sizes for the effect s of time (blue), group (gray). (Bottom) Average change over time in MD-weighted strength for patients with low vs. high WMH volume.

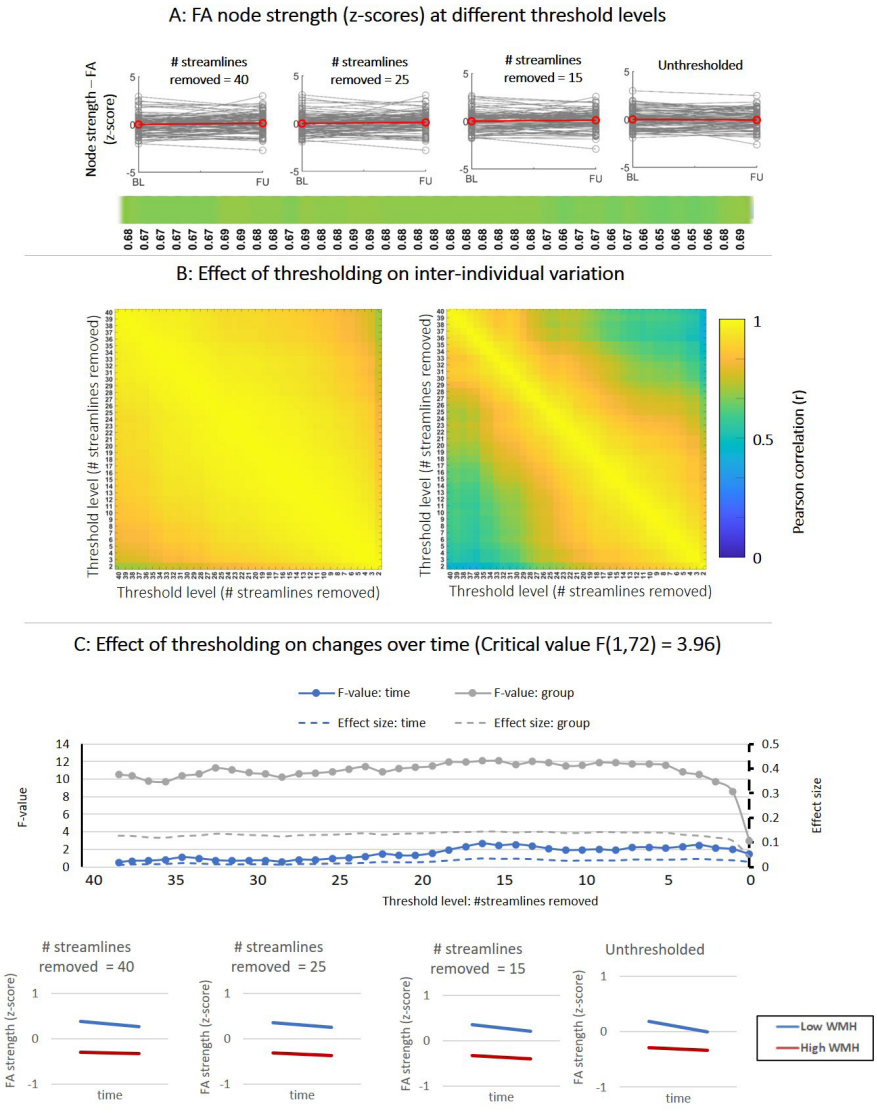


Figure S4: Effect of absolute thresholding on FA-weighted node strength. A: Baseline and follow-up z-scores for different threshold levels. B: (Left) Correlation matrix containing Pearson correlations coefficients between baseline scores of different threshold levels. (Right) Correlations between rate of change over time obtained at different threshold levels. C: Impact of thresholding on the sensitivity to detect changes over time and group-differences in patients stratified by WMH volume. (Top) F-values and effect sizes for the effect s of time (blue), group (gray), interaction term (orange). (Bottom) Average change over time in FA-weighted strength for patients with low vs. high WMH volume.

REFERENCES

Aalten, P., Ramakers, I. H. G. B., Biessels, G. J., de Deyn, P. P., Koek, H. L., OldeRikkert, M. G. M., Oleksik, A. M., Richard, E., Smits, L. L., van Swieten, J. C., Teune, L. K., van der Lugt, A., Barkhof, F., Teunissen, C. E., Rozendaal, N., Verhey, F. R. J., & van der Flier, W. M. (2014). The Dutch Parelsnoer Institute - Neurodegenerative diseases; methods, design and baseline results. BMC Neurology, 14(1), 1–8. https://doi.org/10.1186/s12883-014-0254-4

Andreotti, J., Jann, K., Melie-Garcia, L., Dierks, T., Federspiel, A., & Giezendanner, S. (2014). Repeatability Analysis of Global and Local Metrics of Brain Structural Networks. Brain Connectivity, 4(3), 203–220. https://doi.org/10.1089/brain.2013.0202

Bagarinao, E., Watanabe, H., Maesawa, S., Mori, D., Hara, K., Kawabata, K., Yoneyama, N., Ohdake, R., & Imai, K. (2019). Reorganization of brain networks and its association with general cognitive performance over the adult lifespan. Scientific Reports, June, 1–15. https://doi.org/10.1038/s41598-019-47922-x

Bastiani, M., Shah, N.J., Goebel, R., Roebroeck, A., 2012. Human cortical connectome reconstruction from diffusion weighted MRI: the effect of tractography algorithm. Neuroimage 62 (3), 1732–1749. doi: 10.1016/j.neuroimage.2012.06.002 .

Buchanan CR, Pernet Cyril R., Gorgolewski Krzysztof J., Storkey Amos J., Bastin Mark E.,Test–retest reliability of structural brain networks from diffusion MRI. (2014). NeuroImage (86), 231–243. http://dx.doi.org/10.1016/j.neuroimage.2013.09.054

Buchanan CR, Bastin ME, Ritchie SJ, Liewald DC, Madole JW, Tucker-Drob EM, et al. (2020): The effect of network thresholding and weighting on structural brain networks in the UK Biobank. Neuroimage 211:116443.

Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci. 2009 Mar;10(3):186-98. doi: 10.1038/nrn2575. Epub 2009 Feb 4. Erratum in: Nat Rev Neurosci. 2009 Apr;10(4):312. PMID: 19190637.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284–290

de Brito Robalo, B. M., Vlegels, N., Meier, J., Leemans, A., & Biessels, G. J. (2020). Effect of Fixed-Density Thresholding on Structural Brain Networks : A Demonstration in Cerebral Small Vessel Disease. Brain Connectivity.Apr 2020.121-133.http://doi.org/10.1089/brain.2019.0686., https://doi.org/10.1089/brain.2019.0686

Dimitriadis, S. I., Drakesmith, M., Bells, S., Parker, G. D., Linden, D. E., & Jones, D. K. (2017). Improving the reliability of network metrics in structural brain networks by integrating different network weighting strategies into a single graph. Frontiers in Neuroscience, 11(DEC), 1–17. https://doi.org/10.3389/fnins.2017.00694

Drakesmith, M., Caeyenberghs, K., Dutt, A., Lewis, G., David, A. S., & Jones, D. K. (2015). Overcoming the effects of false positives and threshold bias in graph theoretical analyses of neuroimaging data. NeuroImage, 118, 313–333. https://doi.org/10.1016/j.neuroimage.2015.05.011

Folstein MF, Folstein SE, McHugh PR. "Mini-mental state." J Psychiatr Res. (1975) 12:189–98. 10.1016/0022-3956(75)90026-6

Fornito, A., Zalesky, A., & Breakspear, M. (2013). Graph analysis of the human connectome: Promise,

progress, and pitfalls. NeuroImage, 80, 426–444. https://doi.org/10.1016/j.neuroimage.2013.04.087

Garrison, K. A., Scheinost, D., Finn, E. S., Shen, X., & Constable, R. T. (2015). The (in)stability of functional brain network measures across thresholds. NeuroImage, 118, 651–661. https://doi.org/10.1016/j.neuroimage.2015.05.046

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Van Wedeen, J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. PLoS Biology, 6(7), 1479–1493. https://doi.org/10.1371/journal.pbio.0060159

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V. J., Meuli, R., & Thiran, J. (2007). Mapping Human Whole-Brain Structural Networks with Diffusion MRI. PLoS ONE. 7. https://doi.org/10.1371/journal.pone.0000597

Heinen R, Vlegels N, de Bresser J, et al. The cumulative effect of small vessel disease lesions is reflected in structural brain networks of memory clinic patients. Neuroimage Clin. 2018;19:963-969. Published 2018 Jun 19. doi:10.1016/j.nicl.2018.06.025

Jeurissen, B., Descoteaux, M., Mori, S., & Leemans, A. (2017). Diffusion MRI fiber tractography of the brain. NMR in Biomedicine, July, e3785. https://doi.org/10.1002/nbm.3785

Jeurissen B, Leemans A, Jones DK, Tournier JD, Sijbers J. 2011. Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. Hum Brain Mapp 32: 461–479.

Lawrence, A. J., Tozer, D. J., Stamatakis, E. A., & Markus, H. S. (2018). A comparison of functional and tractography based networks in cerebral small vessel disease. NeuroImage: Clinical, 18(January), 425–432. https://doi.org/10.1016/j.nicl.2018.02.013

Lawrence, A. J., Zeestraten, E. A., Benjamin, P., Lambert, C. P., Morris, R. G., Barrick, T. R., & Markus, H. S. (2018). Longitudinal decline in structural networks predicts dementia in cerebral small vessel disease. Neurology, 90(21), e1898–e1910. https://doi.org/10.1212/WNL.0000000000005551

Lawrence AJ, Chung AW, Morris RG, Markus HS, Barrick TR. (2014). Structural network efficiency is associated with cognitive impairment in small-vessel disease. Neurology 83:304–311.

Leemans A, Jeruissen B, Sijbers J, Jones DK. 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. Proc Int Soc Magn Reson Med 245:3537.

Leemans A, Jones DK. 2009. The B-matrix must be rotated when correcting for subject motion in DTI data. Magn Reson Med 61:1336–1349.

McColgan P, Blom T, Rees G, et al. Stability and sensitivity of structural connectomes: effect of thresholding and filtering and demonstration in neurodegeneration. bioRxiv; 2018. DOI: 10.1101/416826.

Messaritaki, E., Dimitriadis, S. I., & Jones, D. K. (2019). Optimization of graph construction can significantly increase the power of structural brain network studies. NeuroImage, 199(May),495–511. https://doi.org/10.1016/j.neuroimage.2019.05.052

Morris, J.C et al., 1993. The clinical dementia rating (CDR): current vision and scoring rules. Neurology 43, 2412–2414.

Nicols, T.E., Das, S., Eickhoff, S.B., Evans, A.C., Glastard, T.G., Hanke, M., Kriegeskorte, N., Milham, M.P., Poldrack, R.A., Poline, J.-B., Proal, E., Thirion, B., Van Essen, D.C., White, T., Yeo, B.T.T., (2016). Best practices in data analysis and sharing in neuroimaging using MRI. bioRxiv.

Owen, J.P., Ziv, E., Bukshpun, P., Pojman, N., Wakahiro, M., Berman, J.I., Roberts, T.P.L., Friedman, E.J., Sherr, E.H., Mukherjee, P., 2013. Test–retest reliability ofcomputational network measurements derived from the structural connectome of the human brain. Brain Connect. 3, 160–176.

Petersen M. et al., (2020). Network Localisation of White Matter Damage in cerebral Small Vessel Disease, Scientific Reports. 10:9210 | https://doi.org/10.1038/s41598-020-66013-w

Reijmer YD, Leemans A, Brundel M, Kappelle LJ, Biessels GJ. Utrecht Vascular Cognitive Impairment (VCI) Study GroupDisruption of the cerebral white matter network is related to slowing of information processing speed in patients with type 2 diabetes. Diabetes (2013); 62: 2112–5

Roberts JA, Perry A, Roberts G, Mitchell PB, Breakspear M. Consistency-based thresholding of the human connectome. Neuroimage. 2017 Jan 15;145(Pt A):118-129. doi: c10.1016/j.neuroimage.2016.09.053. Epub 2016 Sep 22. PMID: 27666386.

Roine, T., Jeurissen, B., Perrone, D., Aelterman, J., Philips, W., Sijbers, J., & Leemans, A. (2019). Reproducibility and intercorrelation of graph theoretical measures in structural brain connectivity networks. Medical Image Analysis, 52, 56–67. https://doi.org/10.1016/j.media.2018.10.009

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. NeuroImage, 52(3), 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

Sarwar, T., Ramamohanarao, K., & Zalesky, A. (2018). Mapping connectomes with diffusion MRI: deterministic or probabilistic tractography? Magnetic Resonance in Medicine, April 2018, 1368–1384. https://doi.org/10.1002/mrm.27471

Segarra, S., Ribeiro, A., 2014. A stable betweenness centrality measure in networks. IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence, Italy.

Shrout et al., Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420-8. doi: 10.1037//0033-2909.86.2.420. PMID: 18839484.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-berg, H., Bannister, P. R., Luca, M. De, Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. De, Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage. 23, 208–219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Sotiropoulos, S. N., & Zalesky, A. (2017). Building connectomes using diffusion MRI: Why, how and but. NMR in Biomedicine, June 2016, 1–23. https://doi.org/10.1002/nbm.3752

Sporns, O., Tononi, G., & Kötter, R. (2005). The human connectome: A structural description of the human brain. PLoS Computational Biology, 1(4), 0245–0251. https://doi.org/10.1371/journal.pcbi.0010042

Steenwijk MD, Pouwels PJW, Daams M, van Dalen JW, Caan MWA, Richard E, et al. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage Clin. (2013) 3:462–9. doi: 10.1016/j.nicl.2013.10.003

Tax CMW, Otte WM, Viergever MA, Dijkhuizen RM, Leemans A. 2015. REKINDLE: Robust Extraction of Kurtosis INDices with linear estimation. Magn Reson Med 73:794–808.

Tijms, B. M., Wink, A. M., de Haan, W., van der Flier, W. M., Stam, C. J., Scheltens, P., & Barkhof, F. (2013). Alzheimer's disease: connecting findings from graph theoretical studies of brain networks. Neurobiology of Aging, 34(8), 2023–2036. https://doi.org/10.1016/j.neurobiolaging.2013.02.020

Tsai, S. Y. (2018). Reproducibility of structural brain connectivity and network metrics using probabilistic diffusion tractography. Scientific Reports, 8(1), 1–12. https://doi.org/10.1038/s41598-018-29943-0

Tuladhar, anil M., Tay, J., van leijsen, esther, lawrence, andrew J., Wilhelmina Maria van Uden, ingeborg, Bergkamp, M., van der holst, ellen, Kessels, roy P., norris, D., Markus, hugh, & De leeuw, F. (2019).

Structural network changes in cerebral small vessel disease. J Neurol Neurosurg Psychiatry, 0(Run Dmc), 1–8. https://doi.org/10.1136/jnnp-2019-321767

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage, 15(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T. T., & Schmidt, R. (2017). Proportional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. NeuroImage, 152(December 2016), 437–449. https://doi.org/10.1016/j.neuroimage.2017.02.005

van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. In Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2013.09.012

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. NeuroImage, 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

van Wijk, B. C. M., Stam, C. J., & Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. PLoS ONE, 5(10). https://doi.org/10.1371/journal.pone.0013701

Veraart, J., Sijbers, J., Sunaert, S., Leemans, A., & Jeurissen, B. (2013). Weighted linear least squares estimation of diffusion MRI parameters: Strengths, limitations, and pitfalls. NeuroImage, 81, 335–346. https://doi.org/10.1016/j.neuroimage.2013.05.028

Vos SB, Tax CM, Luijten PR, Ourselin S, Leemans A, Froeling M. The importance of correcting for signal drift in diffusion MRI. Magn Reson Med. (2017) Jan;77(1):285-299. doi: 10.1002/mrm.26124.

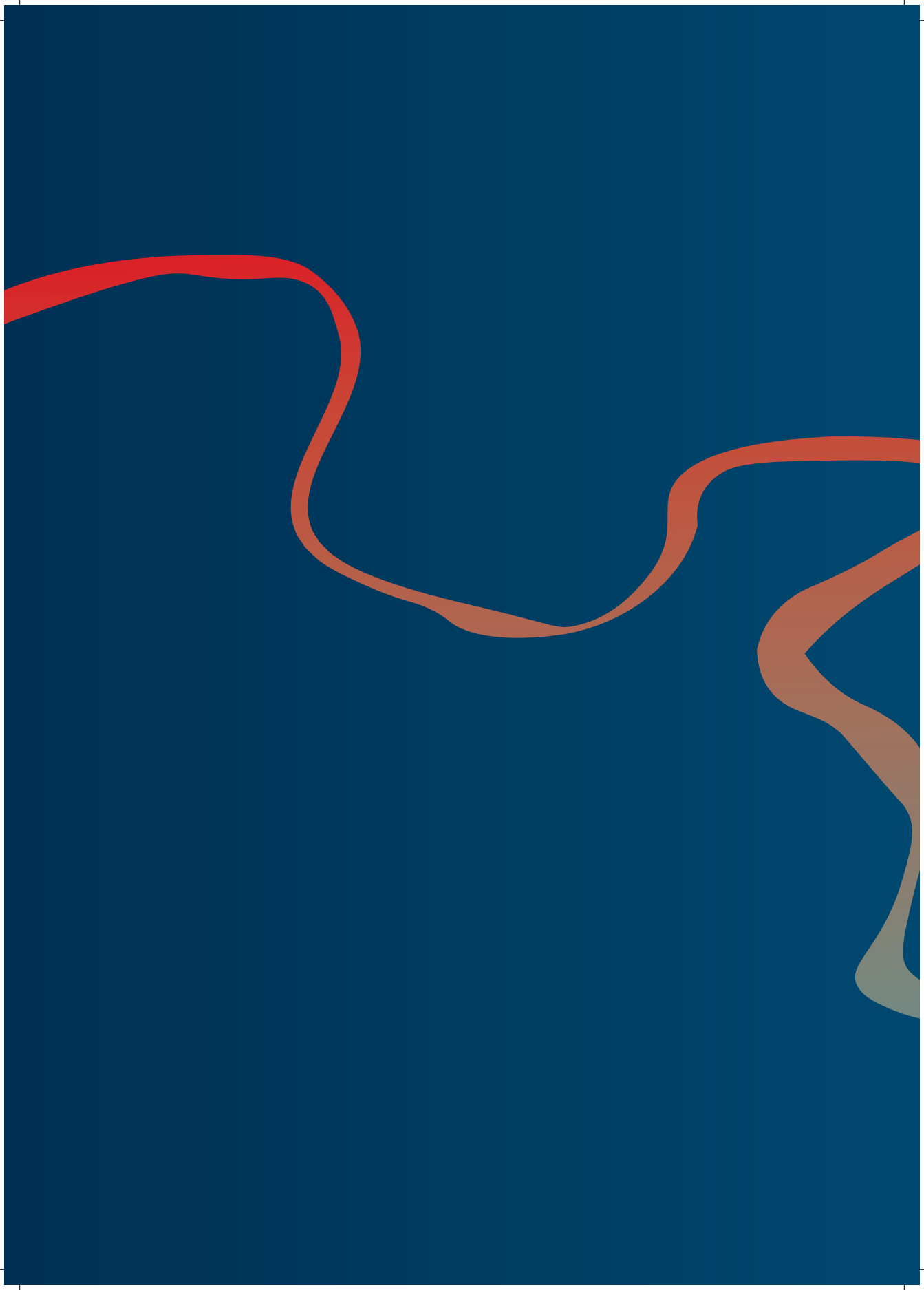
Wang, M., Yuan, Z., & Niu, H. (2019). Reliability evaluation on weighted graph metrics of fNIRS brain networks. 9(5), 832–841. https://doi.org/10.21037/qims.2019.05.08

Welton, T., Kent, D. A., Auer, D. P., & Dineen, R. A. (2015). Reproducibility of Graph-Theoretic Brain Network Metrics: A Systematic Review. Brain Connectivity, 5(4), 193–202. https://doi.org/10.1089/brain.2014.0313

Yuan, J. P., Henje Blom, E., Flynn, T., Chen, Y., Ho, T. C., Connolly, C. G., Dumont Walter, R. A., Yang, T. T., Xu, D., & Tymofiyeva, O. (2019). Test-retest reliability of graph theoretic metrics in adolescent brains. Brain Connectivity, 9(2), 144–154. https://doi.org/10.1089/brain.2018.0580

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., van den Heuvel, M. P., & Breakspear, M. (2016). Connectome sensitivity or specificity: which is more important? NeuroImage, 142, 407–420. https://doi.org/10.1016/j.neuroimage.2016.06.035

Zalesky, A., Fornito, A., Harding, I. H., Cocchi, L., Yücel, M., Pantelis, C., & Bullmore, E. T. (2010). Whole-brain anatomical networks: Does the choice of nodes matter? NeuroImage, 50(3), 970–983. https://doi.org/10.1016/j.neuroimage.2009.12.027.



Chapter 5

Improved sensitivity and precision in multicentre diffusion MRI network analysis using thresholding and harmonization

Bruno M. de Brito Robalo
Alberto de Luca
Christopher Chen
Anna Dewenter
Marco Duering
Saima Hilal
Huiberdina L. Koek
Anna Kopczak
Bonnie Yin Ka Lam
Alexander Leemans
Vincent Mok
Laurien P. Onkenhout
Hilde van den Brink
Geert Jan Biessels

Neuroimage: Clinical 2022, *Accepted*

ABSTRACT

Introduction: To investigate if network thresholding and raw data harmonization improve consistency of diffusion MRI (dMRI)-based brain networks while also increasing precision and sensitivity to detect disease effects in multicentre datasets.

Methods: Brain networks were reconstructed from dMRI of five samples with cerebral small vessel disease (SVD; 629 patients, 166 controls), as a clinically relevant exemplar condition for studies on network integrity. We evaluated consistency of network architecture in age-matched controls, by calculating cross-site differences in connection probability and fractional anisotropy (FA). Subsequently we evaluated precision and sensitivity to disease effects by identifying connections with low FA in sporadic SVD patients relative to controls, using more severely affected patients with a pure form of genetically defined SVD as reference.

Results: In controls, thresholding and harmonization improved consistency of network architecture, minimizing cross-site differences in connection probability and FA. In patients relative to controls, thresholding improved precision to detect disrupted connections by removing false positive connections (precision, before: 0.09-0.19; after: 0.38-0.70). Before harmonization, sensitivity was low within individual sites, with few connections surviving multiple testing correction (k=0-25 connections). Harmonization and pooling improved sensitivity (k=38), while also achieving higher precision when combined with thresholding (0.97).

Conclusions: We demonstrated that network consistency, precision and sensitivity to detect disease effects in SVD are improved by thresholding and harmonization. We recommend introducing these techniques to leverage large existing multicentre datasets to better understand the impact of disease on brain networks.

1 INTRODUCTION

Describing the human brain as a network of nodes (gray matter regions) connected by edges (white matter pathways) can provide powerful insights into neurological disorders (Bullmore & Sporns, 2009; van den Heuvel & Sporns, 2013). Changes in local and global properties of brain networks reconstructed from diffusion MRI (dMRI) have been associated with many disease processes (Pievani et al., 2014). An important issue in this rapidly evolving research field is that network analysis is highly dependent on how the network nodes and connections are defined, particularly when trying to identify specific connections or subnetworks that are disrupted by particular disease. To date, network studies have used various definitions for edges and nodes, different tractography algorithms and edge-weights, resulting in inconsistent reported architectures across datasets (for example, Du et al., 2020; Tuladhar et al., 2017).

Without proper processing, brain networks reconstructed from dMRI are known to contain many false-positive connections which contribute for variability in network architecture (Buchanan et al., 2014). This can be alleviated by thresholding the network to a core structure with more reliable connections, over which eventual disease effects can be assessed with higher confidence (McColgan et al., 2018). A second issue that needs to be considered when detecting network injury patterns in relation to disease is statistical power. To identify specific disrupted connections, methods such as network-based statistics (NBS) are applied to fit a univariate model at every connection and test hypotheses of interest (e.g., differences in fractional anisotropy – FA – between patients and controls at connection level) (Zalesky et al., 2010). Since networks typically contain thousands of connections, multiple tests are performed, often in study populations with a modest sample size, resulting in type I and type II errors. In this scenario, filtering for potentially false-positive connections reduces multiple testing. Moreover, an obvious strategy to improve power is to increase the sample size by pooling existing data from different sites. To date, however, dMRI data pooling has been hampered by acquisition-related differences across sites (Bonilha et al., 2015; Vollmar et al., 2010). As a consequence, network studies of network topology in patients have been predominately single centre with inconsistent recognition of network connections affected by disease (Du et al., 2020; Reijmer et al., 2016; Tuladhar et al., 2017).

Here, we address the challenge of using multicentre dMRI data to study abnormalities in network topology, using cerebral small vessel disease (SVD) as an exemplar condition. In this context SVD is of clear interest because it is considered to affect cognition by disrupting brain connectivity as a result of vascular injury (Frey et al., 2021; Lawrence et al., 2018; Reijmer et al., 2015). However, the exact impact of SVD on network topology is still unknown, with variable results between studies (Du et al., 2020; Tuladhar 2017). This likely relates to the issues raised above, in particular possible confounding by false positive connections and insufficient sample sizes. Recently, in a monocentre setting, we demonstrated the benefits of pre-processing steps such as thresholding to reduce the number of false positive connections and generate more consistent network architectures to assess disease effects (de Brito Robalo et al., 2020). We have also

shown proof of concept of raw data harmonization to minimize acquisition-related differences in multicentre dMRI, at the level of basic diffusion metrics (e.g., FA from the white matter skeleton; de Brito Robalo et al., 2021). In this study, we apply these techniques to higher order diffusion models, to demonstrate the feasibility of multicentre studies of network topology in disease.

We assess the joint application of network thresholding and harmonization to improve sensitivity and precision to detect connections disrupted in SVD. We expect that applying network thresholding leads to more consistent architectures by reducing false positives and improving precision in network analysis, whereas dMRI harmonization improves the sensitivity, by reducing site variability to leverage large multicentre data to increase sample size. We reconstruct dMRI-based networks from four samples with sporadic SVD and one sample with a monogenic form of SVD (Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy – CADASIL, Craggs et al., 2014) and controls. We first evaluate cross-site consistency of network architecture before and after thresholding and harmonization, using age-matched controls from each site. Next, we determine within sites which connections are most affected in patients relative to controls, before and after thresholding and harmonization. To date there is no gold standard for which connections are actually affected by SVD. We therefore use the genetically defined SVD (CADASIL), which is characterized by severe and pure vascular injury, as a reference standard to determine precision and sensitivity for detecting affected connections in the sporadic SVD samples. Finally, we assessed how pooling of scans from all sporadic SVD samples, before and after thresholding and harmonization, improved sensitivity and precision to detect affected connections.

2 METHODS

2.1 Datasets

Four samples of patients with sporadic SVD and one sample with CADASIL were included. Inclusion and exclusion criteria of each cohort are reported in the original studies (summarized below). For the current analysis, we used a harmonized definition for selecting patients and controls from the original cohorts, which was based primarily on the degree of white matter injury, since our objective was to assess benefits of thresholding and harmonization on the detection of injured connections but not their functional impact. Patients had symptomatic SVD defined as a) history of stroke, with a corresponding small subcortical infarct visible on MRI or b) cognitive complaints and presence of white matter hyperintensity (WMH) burden on MRI (Fazekas score ≥ 2, Fazekas et al., 1987). Patients were excluded if they had other major neurological or psychiatric conditions (e.g., multiple sclerosis, epilepsy, Parkinson's disease). Each cohort also included control subjects, with no history of stroke or cognitive complaints and no signs of lacunes on MRI and minimal WMH (Fazekas score 0 or 1). All subjects had T1-weighted and dMRI scans. Characteristics of the study samples (629 patients and 166 controls in total)

are provided in supplementary Table 1. All studies included in this analysis were approved by the ethics committees of the respective institutions and all participants provided written informed consent.

Utrecht: Patients (n = 170) were selected from a memory clinic cohort (Aalten et al., 2014). Controls (n = 46) were recruited from a community-based cohort (Reijmer et al., 2013). MRI scans were acquired on a 3 Tesla Philips scanner (Achieva, Philips, Best, the Netherlands). T1-weighted images had a voxel size of: 1×1×1 mm3, echo time (TE): 4.5 ms and repetition time (TR): 7.9 ms. dMRI data were obtained with a voxel size: 2.5 × 2.5 × 2.5 mm3, TR/TE: 6638/73 ms, 45 diffusion gradients directions with a b-value of 1200 s/mm2, and 1 b = 0 s/mm2 averaged 3 times.

Utrecht 2 (ZOOM): A second dataset from the University Medical Center Utrecht consisted of patients (n = 26) and controls (n = 18) from an ongoing prospective observational cohort study (ZOOM@SVDs, van den Brink et al., 2021). MRI scans were acquired using the same scanner system and acquisition parameters as "Utrecht", albeit with different scanner software versions. Utrecht2 is referred to as "ZOOM" for the remainder of this manuscript.

Hong Kong: Subjects from a community-based cohort that fitted our definition of patients (n = 20) and controls (n = 20) were included (Lam et al., 2019). MRI scans were acquired on a 3 Tesla Philips scanner (Achieva, Philips, Best, the Netherlands). T1-weighted images were obtained with a voxel size: 0.60×1.04×1.04 mm^3, TR/TE: 7.49/3.46 ms and dMRI had a TR/TE: 8944/60 ms, voxel size: 1×1×2 mm3; 32 diffusion gradient directions with b-value 1000 s/mm2 and 1 b = 0 s/mm2.

Singapore: A second community-based cohort was included, with cases (n = 359) and controls (n = 54) fitting our selection criteria (Hilal et al., 2013, 2021). MRI were performed on a 3 Tesla Siemens Magnetom Trio Tim scanner (Siemens Healthineers, Erlangen, Germany). T1-weighted images had TR/TE: 2300/1.9 ms, voxel size: 1×1×1 mm3 and dMRI was acquired with a TR/TE: 6800/85 ms, voxel size: 3.1×3.1×3 mm3; 61 diffusion gradient directions with b-value 1150 s/mm2 and 7 b = 0 s/mm2.

Munich: Patients (n = 54) with CADASIL and controls (n = 28) were selected from a prospective study (Baykara et al., 2016). All MRI scans were acquired on a 3 Tesla Magnetom Verio scanner (Siemens Healthineers, Erlangen, Germany). T1-weighted images were obtained using TR/TE: 2500/4.73 ms, voxel size: 1×1×1 mm3 and dMRI were acquired with a voxel size: 2 × 2 × 2 mm3, TR/TE: 12700/81 ms, 30 diffusion gradient directions with a b-value of 1000 s/mm2, and 1 b = 0 s/mm2.

2.2 Network reconstruction

Diffusion scans were pre-processed using ExploreDTI version 4.8.6 (Leemans et al., 2009). Images were corrected for signal drift (Vos et al., 2017), eddy currents, subject motion with

rotation of the B-matrix (Leemans & Jones, 2009), and susceptibility induced distortions (Veraart et al., 2013). Individual T1-weighted images were resampled to an isotropic resolution of 2 mm^3, then dMRI scans were nonlinearly registered to the resampled T1 images to correct for susceptibility artefacts. Subsequently, the diffusion tensor was estimated with a robust approach (Tax et al., 2015). Deterministic fiber tractography was performed with seed points distributed uniformly throughout the brain, using ExploreDTI. Streamlines were propagated using integration over fiber orientation distributions (FOD), with a step size of 1 mm. FODs were inferred using constrained spherical deconvolution (CSD) with a maximum harmonic order (l-max) of 6 (Jeurissen et al., 2011). Fiber tracking was terminated when streamlines entered a voxel with FOD amplitude < 0.1, or when the deflection was > 45°. Streamlines with a length outside of the bounds 10-500 mm were deemed implausible and excluded. To reconstruct whole brain networks, the Automated Anatomical Labeling (AAL) atlas was used to define 90 cortical and subcortical brain regions that represent network nodes (Tzourio-Mazoyer et al., 2002). Networks were reconstructed by combining the segmented AAL regions with the tractography data. Two nodes were considered connected if they contained the end-points of at least one streamline, resulting in 90 × 90 binary connectivity matrices. The connectivity matrices were also weighted by fractional anisotropy (FA) to obtain 90 x 90 FA-weighted connectivity matrices.

2.3 Analysis

2.3.1 Cross-site consistency of network architecture

We assess whether harmonization and thresholding improve the consistency of network architecture across sites. For this objective we focused on networks of controls, which were age-matched across sites (n =15 for each site). The most straightforward definitions of network architecture are the presence or absence of connections (i.e., the binary structure) and their corresponding weights (e.g., FA). These characteristics constitute the basis of all network properties from which all network metrics are derived (Messaritaki et al., 2019). To evaluate consistency of the binary structure, we first constructed a group-level probability matrix for each site by averaging the binary connectivity matrices of all subjects. In this probability matrix, each entry represents the probability a connection being reconstructed in that group. For example, a connection with a probability of 0.5 is detected in 50 % of subjects. Second, we defined consistency between two sites as the average relative difference (in %) between the group-level probability matrices.

$$difference = 100\% \times \frac{|Site1 - Site2|}{\frac{Site1 + Site2}{2}} \qquad (1)$$

Note that the lower the difference between sites the higher the consistency in connection probability. To evaluate consistency of FA-weights, we also constructed a group group-level matrix containing the average FA of connections reconstructed within each group. We computed the consistency in FA as the difference between the group-level FA-matrices of different sites as by the formula above. We determined the consistency before and after thresholding and harmoni-

zation, with each method being applied separately and in combination.

Thresholding was performed on the group-level probability matrix by removing connections with low probability until a fixed network density of 15 % was achieved (de Reus & van den Heuvel, 2013; Roberts et al., 2017). This thresholding approach removes the most improbable connections within each site, while also ensuring that networks with equal densities are achieved across all sites.

Harmonization was performed by scaling the dMRI signal off 15 age and sex-matched controls of all sites to an arbitrary reference (Utrecht) using rotation invariant spherical harmonic (RISH) features (Karayumak et al., 2019). Both our previous work (de Brito Robalo et al., 2021) and other method papers (Mirzaalian et al., 2015; Karayumak et al., 2019) have shown that 15 to 20 controls is sufficient to calculate average RISH features that capture group properties of each site, rather than individual characteristics, to perform effective harmonization. This harmonization method is applied to the raw dMRI signal to remove acquisition-related differences while preserving biological effects and between-subject variation.

2.3.2 Sensitivity and precision to disease effects in SVD

Here we assess if thresholding and harmonization improve precision and sensitivity to detect connections affected in patients with SVD. Clearly there is no established ground-truth, apart from histology, on which tracts are preferentially affected by SVD and which are relatively spared (Craggs et al., 2014). Moreover, substantial inter-individual variation is likely to occur, according to lesion burden and location. Yet, we required a reference standard to provide a basis to test the validity of our approach. For this we used a sample of patients with a monogenic form of SVD (CADASIL) with a high burden of disease, as a pure form of SVD to which we could compare the findings in the other samples with sporadic SVD (Di Donato et al., 2017). To identify connections most affected in SVD, network-based statistics adjusted for age and sex was performed to compare the FA of each connection between patients and controls. Specifically, two-sample t-tests were performed for each connection and False Discorvery Rate (FDR) correction with 50 thousand permutations was used to account for multiple comparisons. Connections with large effect sizes (Cohen's $d > 0.8$) relative to controls that survive network thresholding and FDR in the CADASIL cohort were used as reference standard of SVD-related patterns of network injury.

To estimate precision and sensitivity within each sporadic SVD cohort, NBS with the same parameters as described above was performed to detect connections strongly affected in patients relative to controls. These affected connections were compared to those found in the CADASIL reference. Precision was defined as the ratio between true positives (TP; i.e., connections also affected in the reference standard) and the total number of connections detected:

$$precision = \frac{TP}{TP + FP} \qquad (2)$$

Here, false positives (FP) are connections detected in sporadic SVD but not in the reference. Precision was estimated for the pooled dataset and all sporadic SVD cohort, before and after harmonization and thresholding. The number of connections with large effect size that survive thresholding and FDR-correction was used as an indicator of sensitivity.

3 RESULTS

3.1 Cross-site consistency of network architecture

Figure 1 shows group-level connectivity matrices containing connection probability for age-matched controls of each site, before and after harmonization and thresholding. Before thresholding and harmonization, networks contained many connections with low probability (i.e., that were only reconstructed in a small proportion of subjects) (Fig. 1A). These low-probability connections were inconsistent and were reconstructed in different network locations, resulting in large differences in probability across sites (relative cross-site difference: 34.7% - 46.8%, right panel). Harmonization alone did not change the probability of connections nor the consistency across sites (Figure 1B). This implies that on average the same tracts were reconstructed before and after harmonization. When comparing a fibre bundle (e.g., corpus callosum) of one subject before and after harmonization, the same streamlines were reconstructed (see Supplementary Figure S2). Thresholding had a significant impact on the connection probability and their consistency across sites since it removed low probability connections and retained connections with high probability (Fig. 1C and 1D). Since thresholding is applied at the higher level (on the connectivity matrix) it did not directly affect tractography results but rather determined which fibre bundles are considered for the network when defining nodes and edges. This resulted in more consistent network architectures across sites, with high-probability connections detected at the same network locations across sites (relative cross-site difference: 4.5% - 7.8%).

Figure 2 shows group-level FA-weighted connectivity matrices within each site, before and after harmonization and thresholding. FA values varied substantially across sites before harmonization, particularly for Hong Kong (relative cross-site difference up to 36.9%, Fig.2A). After harmonization, cross-site differences in FA across were minimized, even though the network still contained many low-probability connections (relative cross-site difference in FA: 10.8% - 20.5%, Fig. 2B). At the tract level for example in the corpus callosum) it is also clear that FA values change along the tract after harmonization. Thresholding alone did not remove differences in FA across sites, because even though the connections retained were more consistent in probability, the weights of these connections still differed across sites due to scanner-related differences (Fig. 2C). When applied together, thresholding and harmonization produced networks that only contained high-probability connections with minimal differences in FA across sites (Fig. 2D). In conclusion harmonization produced more similar FA values across sites without

changing the reconstruction of tracts.

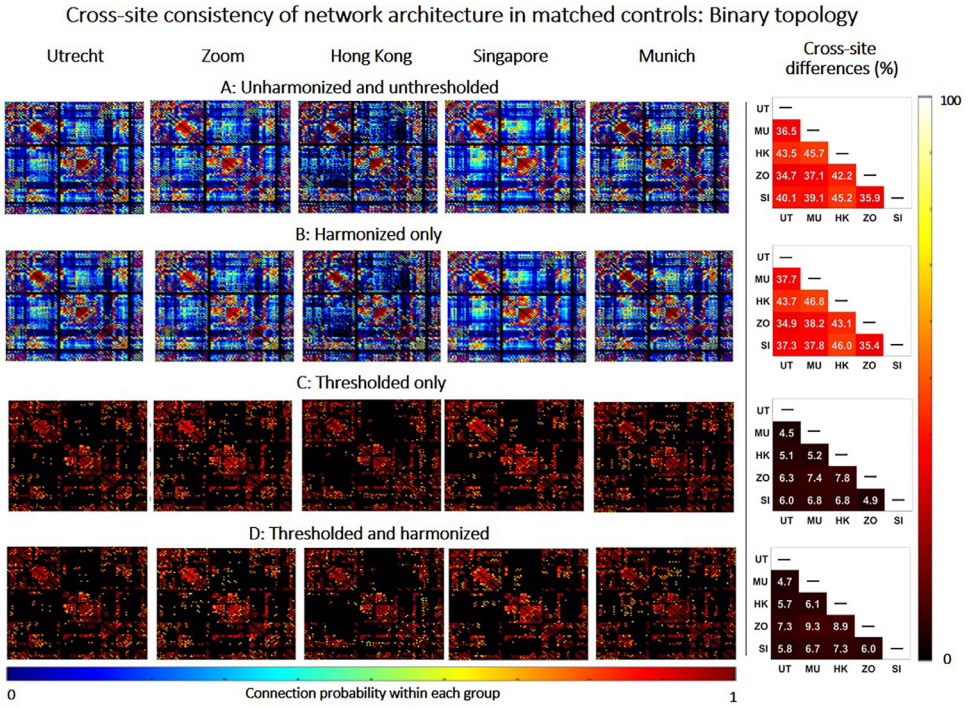


Figure 1: Consistency of network architecture (binary topology) between matched controls of different sites. Left: Connectivity matrices containing the probability of connections within each site (e.g., a connection with a probability of 0.2 is only detected in 20% of subjects of that group). The black background in the connectivity matrix means no connection is found between two nodes. Right: Relative difference between connectivity matrices of different sites. Results are displayed before harmonization and thresholding (A), after harmonization only (B), after thresholding only (C) and after combining the thresholding and harmonization (D).

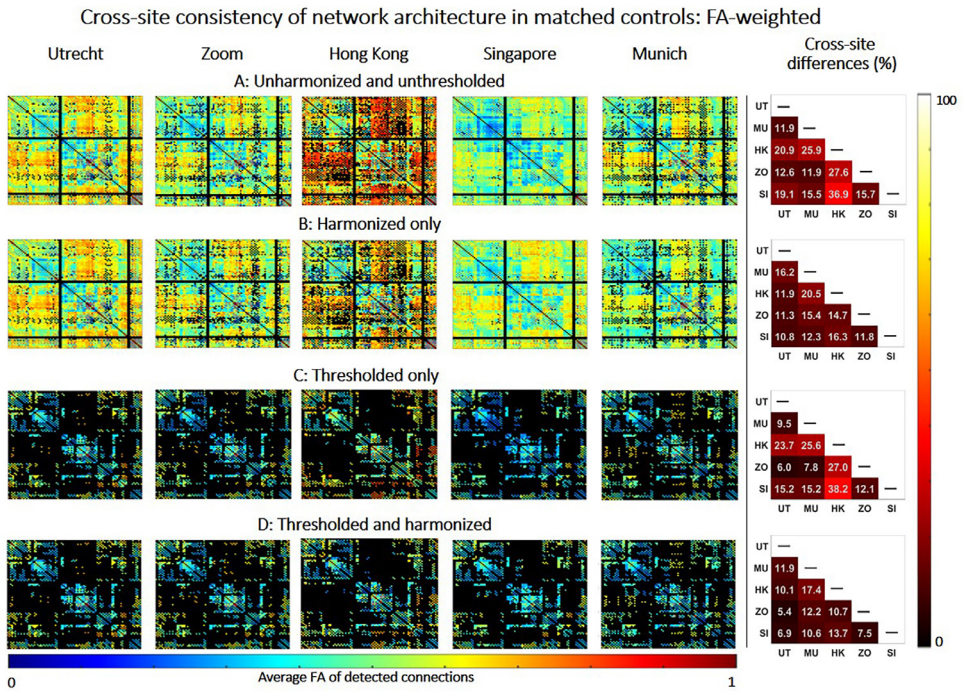


Figure 2: Consistency of network architecture (FA weight) between matched controls of different sites. Left: Connectivity matrix containing the average FA of connections detected within each site. The black background in the connectivity matrix means no connection is found between two nodes. Right: Relative difference between connectivity matrices of different sites. Results are displayed between before harmonization and thresholding (A), after harmonization only (B), after thresholding only (C) and after combining thresholding and harmonization (D).

3.2 Sensitivity and precision to disease effects in SVD

Figure 3 describes patterns of connections affected in patients relative to controls, with results shown for individual samples with sporadic SVD (A), pooled data of all sporadic SVD samples before (B) and after harmonization (C) as well as for the reference CADASIL sample (D). For individual sites with sporadic SVD, different disease burdens were observed across sites -reflecting differences in patient populations- as indicated by the range of effect sizes between patients and controls. Utrecht and Zoom had the most connections with large effect sizes, whereas Singapore and Hong Kong had the lowest disease burden (Fig. 3A, second row). However, in all sites the majority of strongly affected connections were likely spurious as they were only present in few patients and controls (i.e., connections with low probability at group level). When unthresholded connections were compared to those found in the reference CADASIL sample, a large proportion were false positives, which resulted in low precision within sites before thresholding (precision: 0.08 - 0.19, Table 1). Thresholding drastically reduced the number of false positives and improved precision within sites, relative to the reference (precision: 0.38 – 0.64,

Fig. 3A, third row). However, only a small proportion of connections retained after thresholding also remained significant after FDR-correction (25 connections in the Utrecht cohort), indicating low sensitivity within individual sites (Fig. 3A, fourth row).

Regarding the pooled data before harmonization, effect sizes were inflated due to scanner-related differences in FA, resulting in more than 2/3 of all tested connections having large effect sizes (2271 connections with effect size >0.8, Fig. 3B, first and second row). Moreover, the majority of these large effect sizes were found in low-probability connections, which resulted in low precision before thresholding, relative to the reference standard (precision: 0.06, Table 1). After thresholding, low probability connections were removed, resulting in improved precision (precision: 0.70, Table 1). Compared to individual sites, more connections survived FDR-correction in the non-harmonized pooled data (164 connections, Fig. 3B, fourth row), due to the larger sample size. After harmonization, the effect sizes observed in the pooled data were no longer inflated by acquisition-related differences and were closer to the range of effect sizes observed in individual samples with sporadic SVD (Fig.1C, first and second row). Thresholding removed low probability connections, reducing the rate of false positives and improving precision (0.85). Sensitivity was apparently also improved since more connections survived FDR, as compared to individual sites (k = 38, Fig. 3C, fourth row).

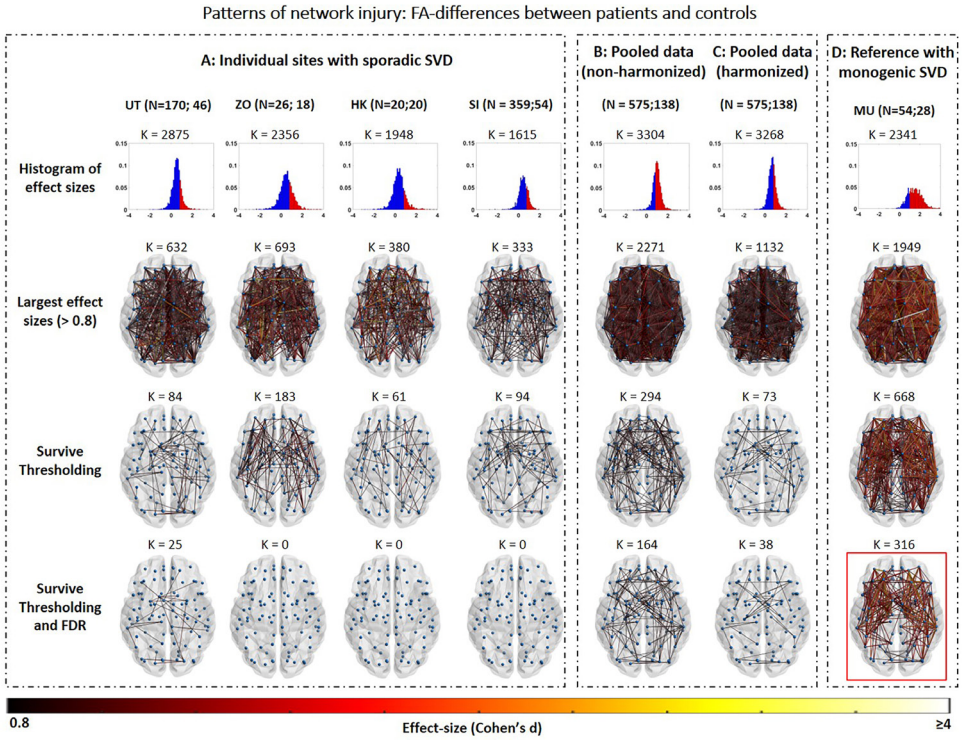


Figure 3: Patterns of network injury within sites with sporadic SVD versus controls (A), pooled data before (B) and after (C) harmonization as well as a reference cohort of CADASIL patients versus controls (D). N=x;y indicates numbers of patients and controls per site or in the pooled cohort. K represents the number of connections tested in the histogram or shown in the brain network map. First row: Histogram of effect sizes, showing Cohen's d for all connections tested between patients and controls. Red area represents effect sizes >0.8. Second row: Connections with large effect sizes (> 0.8) for patients compared to controls. Third row: Connections with large effect sizes that survive network thresholding. Third row: Connections with large effect sizes that survive thresholding and FDR correction. After each step the patterns of connections of each site were compared with the thresholded and FDR-corrected CADASIL cohort shown in red (bottom right). The ratio of true positives, false positives and precision relative to the CADASIL cohort were calculated and shown in Table 1.

4 DISCUSSION

In a first study of its kind, using multicentre data of patients with SVD as an exemplar condition, we demonstrated that combining network thresholding and RISH harmonization improve consistency of dMRI-based brain networks across sites, while also increasing precision and sensitivity to detect connections affected in disease. Thresholding removed low probability connections improving the precision to detect connections most strongly affected within each cohort with sporadic SVD. Yet, individual sites showed low sensitivity to detect strongly affected connections, likely due to modest sample sizes, with only a small number of connections surviving FDR correction. Harmonization minimized cross-site differences in FA and enabled data pooling to achieve a large sample size and improve the sensitivity to detect strongly affected connections. When applied together, these two methods generated networks with more consistent sets of connections and more similar FA-weights across sites, while also achieving higher precision than individual sites.

Thresholding and Harmonization improve network consistency

Preceding our work, several studies have analysed brain networks of older adults and the impact of SVD on global and local network properties as well specific patterns of injury (Du et al., 2020; Lawrence et al., 2018; Tuladhar et al., 2017). However, a frequently reported limitation in those studies is the low generalizability of results across datasets due to sample size or differences in network architecture (Smith et al., 2019). Our results indeed demonstrate that the network architecture before thresholding can differ substantially across age-matched controls with minimal burden of white matter injury. Within different sites, networks contained many unreliable tracts which were only reconstructed in a small proportion of subjects. These inconsistent tracts are likely a byproduct of cumulative errors during network reconstruction steps, especially in fiber tractography (de Reus & van den Heuvel, 2013). These tracts had a wide-range of lengths (10-300 mm), indicating that thresholding was not biased to only short or long-range connections (see Supplementary Figure S1). While no ground-truth brain network has yet been established, it is well accepted that connections that describe underlying white matter pathways should be detected consistently across subjects, with small degree of interindividual variation (Roberts et al., 2017). As shown by our current results as well as previous work, introducing thresholding as a step during network reconstruction removes low probability tracts and generates networks with more consistent architectures across datasets (Buchanan et al., 2020; Kurokawa et al., 2021; Messaritaki et al., 2019). We add to those results by showing that thresholding improves cross-site consistency in elderly subjects. We note that network thresholding is a higher-level step, since it is applied on the connectivity matrices and not on tractography itself. Therefore, it does not directly change the tractography results, but rather determines which edges of the network are more consistently observed across subjects. However the concept of thresholding (i.e., filtering noisy data) could also extend to other network reconstruction pipelines based on probabilistic tractography (Roberts et al., 2017). In this scenario, thresholding could be coupled with other established filtering strategies in probabilistic tractography (SIFT2, Smith et al.,

2015). Of note, the same principles apply to functional connectomes where thresholding has been shown to increase edge-consistency across participants (Váša et al., 2018).

A possible concern about thresholding is that removing connections from the network can be detrimental to the sensitivity to disease effects (van Wijk et al., 2010). Although we cannot formally test sensitivity, because we have no standard of which connections were truly affected in each of the samples with sporadic SVD, it is arguably desirable to establish disease effects over a more consistent network architecture, less affected by spurious tracts (de Brito Robalo et al., 2020). In this work we applied a thresholding method that not only removes low probability connections across subjects, but also ensures that fixed network density is achieved across sites (van den Heuvel & Sporns, 2013). The stringency of thresholding remains arbitrary, but our own work and previous studies have shown that mid-range densities (≈ 15%) reduces the number of false positives connections in networks reconstructed with deterministic fibre tractography, while maintaining the well-known small-world structure of brain networks and sensitivity to disease effects (Hagmann et al., 2008; Robalo et al., 2020; Roberts et al., 2017).

Our results also demonstrated that dMRI harmonization improves the consistency of connection weights across sites. Before harmonization, FA of reconstructed connections varied substantially between matched controls of different sites due to acquisition-related differences in the diffusion signal. Groups of healthy subjects that are matched for demographics should have similar diffusion profiles and therefore scalar diffusion metrics used for network weighting should not significantly differ across sites (Mirzaalian et al., 2015). RISH harmonization removed differences in the diffusion signal and produced networks with more similar FA across sites. The harmonization results shown in this study should be generalizable in other studies since it is applied at the beginning of the processing pipeline, prior to other methodological considerations. (e.g., FA, MD). Since the RISH method only requires a subset of matched subjects from each site as training subjects, it can be generalized to any multicentre cohort as long as training data is available. Importantly, researchers should be aware that subject matching is a challenging step, especially when dealing with elderly subjects or subjects with brain lesions. In those cases we recommend ensuring that brain lesions in the training data are comparable across sites, both in terms of location and severity.

While alternative methods for harmonization exist (e.g. statistical harmonization with ComBat, Fortin et al., 2017, mega and meta analysis), it remains unclear how to couple them with connectivity analysis where several network weights can be used and different graph metrics are derived for further analysis. Contrary to meta and mega-analysis which combine statistical results from already processed data (e.g., effect sizes), the RISH method aims at creating a single large dataset, allowing any subsequent analysis (such as network analysis or FADTTS). When the data is available, RISH harmonization should be prioritized over meta and mega-analysis. Here, we used FA-weighted networks, since this metric has been widely used as outcome in network studies of SVD to examine contrasts between patients and controls or association with cognition (Heinen et al., 2018; Reijmer et al., 2016). A recent report (Schilling et al., 2021)

suggests that variability in scanner hardware and acquisition parameters is the largest source of uncertainty when performing fiber tractography with multi-site data. In our study, we did not observe large variability in brain networks before and after harmonization, which mainly impacted dMRI metrics of the connections, rather than their probability. This might imply that connection probability in brain connectivity is overall resistant to inter-scanner difference, especially when a relatively coarse-grained atlas is used to define the nodes (e.g. AAL atlas). On the other hand, data used in this work only included modest diffusion weightings, but the use of larger diffusion weightings – which is reccomended for accurate fiber tractography (Tournier et al., 2008) – might further exacerbate hardware-driven differences. Overall, when harmonization was combined with thresholding, the resulting networks had more consistent connections across subjects and across sites and more similar FA-weights. Overall, when harmonization was combined with thresholding, the resulting networks had more consistent connections across subjects and across sites and more similar FA-weights. A thresholded and harmonized network is thus a more reliable basis from which patterns of injury in SVD can be identified.

Thresholding and Harmonization improve sensitivity and precision to effects of interest

In the second part of this study, we examined the impact of thresholding, harmonization and pooling on the sensitivity and precision to detect connections strongly affected in SVD. Our findings within sporadic SVD samples show that before thresholding there is low precision to detect relevant connections compared to the reference standard. This is in part due to the fact that large effect sizes were detected in low probability connections, which were absent in a large proportion of subjects, and likely represent artefacts. For these low probability connections, the number of data points used for the t-tests in the NBS pipeline is even smaller than the already small sample size within sites (number of patients and controls). This resulted in many outlier connections showing large effect sizes (> 0.8). Thresholding removed a large proportion of false positive connections, ensuring that all large effect sizes were identified only for connections that were present in a large proportion of subjects (i.e., the actual sample size used for the t-test was closer to the group sample size). This resulted in higher precision to detect strongly affected connections. This finding is in line with previous research with healthy subjects from the UK biobank, which showed that connections retained in the network after thresholding are more strongly associated with biological effects of interest, such as age (Buchanan et al., 2020).

Within each site, sensitivity was very low since few connections remained significant after FDR-correction. Significant connections were only retained in the Utrecht cohort, which had relatively high disease burden and a substantial sample size. In all other sites, no significantly different connection between patients and controls remained after FDR, likely due to limited sample size. For example, the Zoom Cohort had effect sizes relatively similar to Utrecht. For example, the Zoom Cohort had relatively similar effect sizes as Utrecht, but a limited sample small size to detect significant connections. On the other hand, Singapore had the largest sample size from all sites but since patients were community-duelling individuals with lower disease burden and more subtle effect sizes, but no connection survived FDR. So far, most network

analysis studies in SVD employed limited sample sizes, with the exception of a recent work with 930 subjects that identified key connections disrupted in SVD, comprising interhemispheric connections as well as connections between subcortical and frontal brain regions (Petersen et al., 2020).

An application of network localisation of SVD-related injury could be to help understanding how diffuse and/or focal damage in certain brain areas affects cognitive function (Biesbroek et al., 2018). If these findings are to be translated in more applied settings, they should be validated with large multicentre datasets and generalized across centers. With our pooled analysis we demonstrated that combining multicentre data to achieve a larger sample size can improve the sensitivity to detect connections affected in SVD, which would not have been possible using single center data. Importantly, our results clearly show that this should only be performed after harmonization to avoid introducing bias in effect sizes, in analogy to previous studies on diffusion tensor imaging metrics (Cetin-Karayumak et al., 2020; de Brito Robalo et al., 2021). When the data was pooled without harmonization, effect sizes were much larger than those observed within individual sites, which is likely erroneous. Consequently, more than 2/3 of all the tested connections had an effect size larger than 0.8, which originates from acquisition-related differences in FA across sites, combined with imbalance in proportions of controls and patients in the samples. Even though more "affected connections" were detected, these were mostly false positives, resulting in low precision. Of note, without harmonization, the opposite scenario could also be possible, i.e., where pooling data results in a loss of effect size rather than inflation, depending on patient and control sampling and whether scanner effects produce higher or lower FA values for one site, relative to another. To avoid these detrimental scenarios, it is crucial that harmonization is performed prior to data pooling to ensure that the detected differences are not driven by acquisition-related differences across sites but by the increased power of a larger sample size. In summary, when applied together, thresholding, harmonization and pooling improved sensitivity to detect connections affected in sporadic SVD with higher precision than individual sites.

Limitations and future outlook

A limitation of our study is the lack of a ground-truth for patterns of injured connections in SVD. By defining the CADASIL sample as model for patterns of network injury in SVD, we created a reference standard to compare the findings in patients with sporadic SVD from each site to. We realize that patterns of injury observed in CADASIL do not represent an actual ground-truth of network injury in sporadic SVD, but rather represent a typical example of most severely affected tracts to support our analyses. Using this reference standard, we focused on improving processing steps to identify network injury. Of note, at this stage we did not aim to obtain novel insight into actual SVD disease effects. Accordingly, we did not harmonize patient selection for particular disease stages, risk profiles, or outcomes. Our selection was primarily based on white matter lesion burden. Future studies could use the techniques demonstrated here to further characterize the actual patterns of disrupted connections in patients with SVD and their functional impact.

This should also entail more rigorous selection of patients and controls, considering parameters such as risk factors, comorbidities or clinical diagnosis (e.g., mild cognitive impairment; dementia). Furthermore, recently developed techniques such as machine learning and classification algorithms could benefit from large sample sizes gained by harmonizing multicentre data. In future studies it could be of interest to use classifiers to determine which patients are likely to have cognitive decline over time based on network metrics or which white matter tracts can predict change in specific cognitive domains. For such studies to be feasible it is important to have a large test sample to train the classifier and to perform validation.

Our study was also limited to datasets with lower quality (e.g., lower angular resolution) as compared to other connectivity studies in healthy young adults (Sotiropoulos et al., 2013), but on the other hand reflects the reality in a clinical setting. Another point to consider from a technical perspective, is that network reconstruction involves many methodological choices which could not all be tested in this study, and are topics of debate (e.g., registration methods, tractography algorithm, parcellation schemes, Yeh et al., 2021). Finally, we examined network consistency based on primary features of network architecture (e.g., connection probability and connection weight), but topological metrics derived from graph theory should also be investigated in future analyses (Bullmore & Sporns, 2009).

4 CONCLUSION

Using multicentre dataset of patients with SVD, we demonstrated that sensitivity and precision to detect disease effects in specific network connections can be improved by thresholding, harmonization and pooling. Future studies that investigate network localisation of SVD-related injury, or other diseases, and associations with brain function should introduce thresholding as part of the processing pipeline, to ensure that disease effects are evaluated over a network with a more consistent architectures and fewer positives. Furthermore, since effect sizes are likely to be subtle, large sample sizes should be used to study these effects. We provide proof of concept that harmonization facilitates pooling of large multicentre datasets to achieve this goal.

ACKNOWLEDGEMENTS

We acknowledge and thank all patients and controls for study participation. We thank members of the Utrecht Vascular Cognitive Impairment (VCI) Study group. We also thank Mathias Hübner for technical assistance and Angelika Doerr as study nurse for her help in the VASCAMY study in Munich.

SUPPLEMENTARY INFORMATION

Table 1: Demographics and imaging parameters of the study samples.

	Utrecht 1		Utrecht 2 (ZOOM)		Hong Kong		Singapore		Munich	
	Controls (N = 46)	Patients (N = 170)	Controls (N = 18)	Patients (N = 26)	Controls (N = 20)	Patients (N = 20)	Controls (N = 54)	Patients (N = 359)	Controls (N = 28)	Patients (N = 54)
Demographics										
Age, years	71.2 ± 4.8	74.9 ± 9.0	62.4 ± 6.9	64.8 ± 6.8	69.2 ± 3.4	74.1 ± 3.3	66.6 ± 4.8	71.3 ± 6.0	71.8 ± 7.4	56.0 ± 7.1
Male sex (%)	28 (61)	99 (59)	10 (55)	18 (69)	10 (50)	10 (50)	31 (52)	166 (46)	16 (57)	20 (37)
MRI markers										
WMH volume	0.3 [0.1, 0.6]	1.2 [0.5, 2.7]	0.03 [0.02, 0.08]	0.7 [0.3, 1.3]	0.1 [0.05, 0.1]	0.6 [0.3, 1.1]	0.04 [0.02, 0.09]	0.8 [0.08, 1.16]	0.2 [0.1 0.7]	6.2 [3.5, 10]
WMH (Fazekas)	0 [0, 1]	2 [1, 2]	1 [1,1]	2 [2,3]	0.5 [0, 1]	2 [2, 3]	1 [1,1]	2 [1,3]	0 [0, 1]	3 [2, 3]
Lacunes (% present)	0 (0)	69 (36)	0(0)	12 (46)	0 (0)	-	0 (0)	118 (33)	0 (0)	40 (74)
Imaging parameters										
Scanner	3T Philips Achieva		3T Philips Achieva		3T Philips Achieva		3T Siemens Magnetom Trio, Tim		3T Siemens Verio	
Software version	R3.1		MR Release 5.6.0		MR Release 5.1		syngo MR B19		syngo MR B19	
Voxel size (mm^3)	2.5 × 2.5 × 2.5		2.5 × 2.5 × 2.5		1 × 1 × 2		3.1 × 3.1 × 3.0		2 × 2 × 2	
b-value (s/mm^2)	1200		1200		1000		1150		1000	
# of directions	45		45		32		61		30	

Data presented as mean ± SD, number (percentages) or median [interquartile range]; SVD = small vessel disease; CADASIL = cerebral autosomal-dominant arteriopathy with subcortical infarcts and leukoencephalopathy; WMH = white matter hyperintensity (normalized to % of intracranial volume).

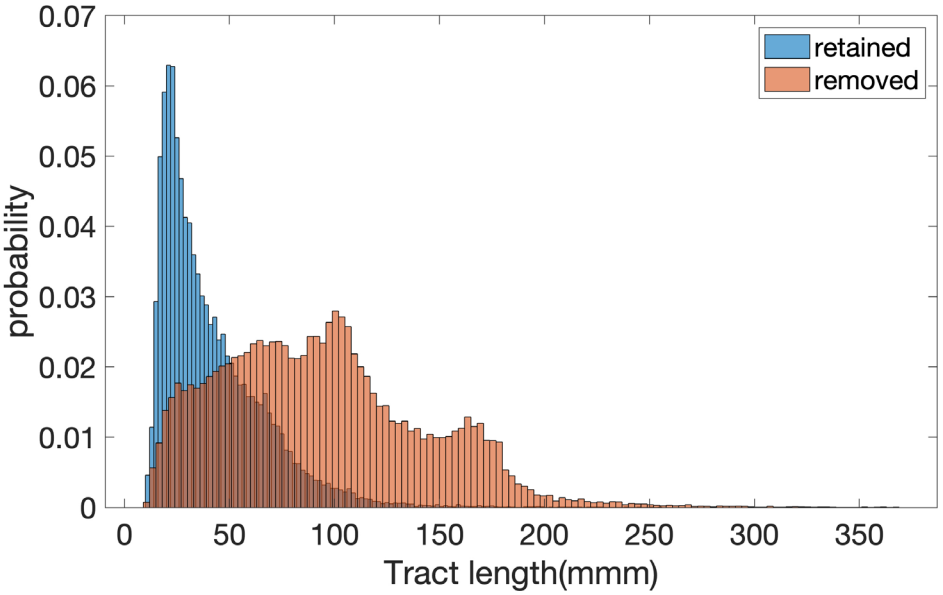


Figure S1: Average tract length distribution of connections retained and connections removed from the network during thresholding.

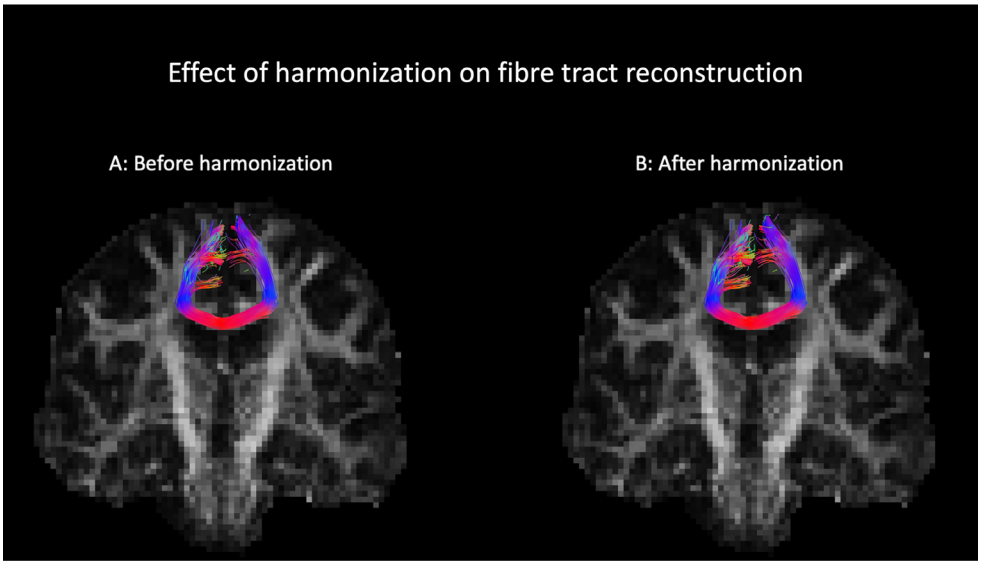


Figure S2: Fiber tractography of the corpus callosum before (A) and after (B) harmonization of an example control subject. The figure shows that harmonization does not change the reconstruction of fiber tracts, nor their pathways. Color coding: red (left-right); blue (superior-inferior); green (anterior-posterior). This supports our main results in Figure 1 where we observed that harmonization alone does not affect the probability of a connection being reconstructed.

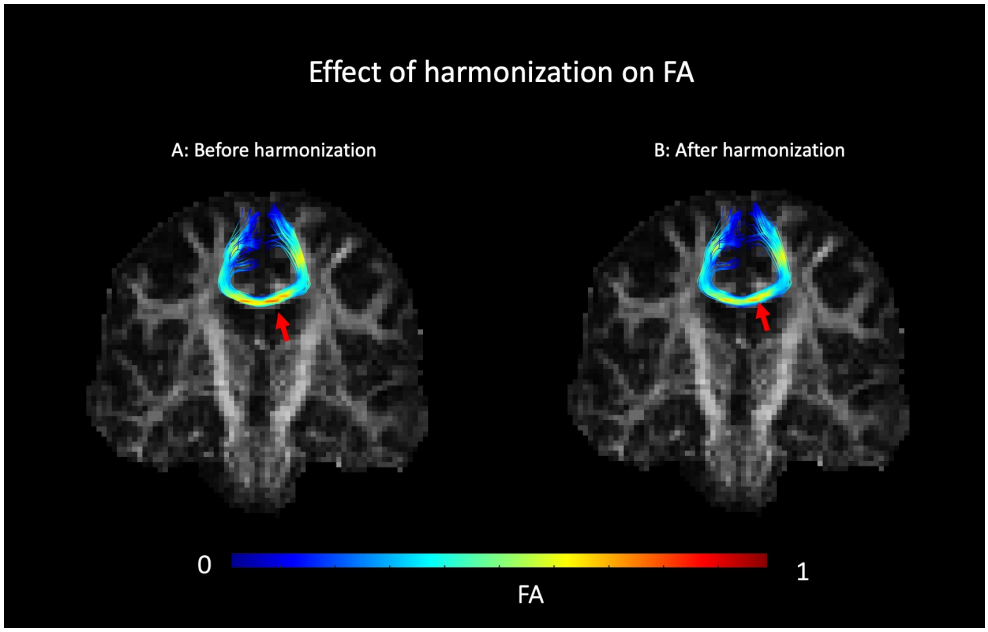


Figure S3: Fiber tractography of the corpus callosum before (A) and after (B) harmonization of an example control subject, color-coded by FA. The arrows indicate regions along the tract where FA changes after harmonization. This is in line with our main results in Figure 2, where we observed that the FA-weighted connectivity matrix is scaled after harmonization and becomes more similar across sites.

5 REFERENCES

Aalten, P., Ramakers, I. H. G. B., Biessels, G. J., de Deyn, P. P., Koek, H. L., OldeRikkert, M. G. M., Oleksik, A. M., Richard, E., Smits, L. L., van Swieten, J. C., Teune, L. K., van der Lugt, A., Barkhof, F., Teunissen, C. E., Rozendaal, N., Verhey, F. R. J., & van der Flier, W. M. (2014). The Dutch Parelsnoer Institute - Neurodegenerative diseases; methods, design and baseline results. BMC Neurology, 14(1), 1–8. https://doi.org/10.1186/s12883-014-0254-4

Baykara, E., Gesierich, B., Adam, R., Tuladhar, A. M., Biesbroek, J. M., Koek, H. L., Ropele, S., Jouvent, E., Chabriat, H., Ertl-Wagner, B., Ewers, M., Schmidt, R., de Leeuw, F. E., Biessels, G. J., Dichgans, M., & Duering, M. (2016). A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. Annals of Neurology, 80(4), 581–592. https://doi.org/10.1002/ana.24758

Biesbroek, J. M., Leemans, A., Den Bakker, H., Duering, M., Gesierich, B., Koek, H. L., Van Den Berg, E., Postma, A., & Biessels, G. J. (2018). Microstructure of Strategic White Matter Tracts and Cognition in Memory Clinic Patients with Vascular Brain Injury. Dementia and Geriatric Cognitive Disorders, 44(5–6), 268–282. https://doi.org/10.1159/000485376

Bonilha, L., Gleichgerrcht, E., Fridriksson, J., Breedlove, J. L., Rorden, C., Nesland, T., Paulus, W., Helms, G., & Focke, N. K. (2015). Reproducibility of the structural brain connectome derived from diffusion tensor imaging. PLoS ONE, 10(9), 1–17. https://doi.org/10.1371/journal.pone.0135247

Brink, H. Van Den, Kopczak, A., Arts, T., Onkenhout, L., Siero, J. C. W., Zwanenburg, J. J. M., Duering, M., Blair, G. W., Doubal, F. N., Stringer, M. S., Thrippleton, M. J., Kuijf, H. J., Luca, A. De, Hendrikse, J., Wardlaw, J. M., Dichgans, M., & Jan, G. (2021). Cerebral Circulation - Cognition and Behavior Zooming in on cerebral small vessel function in small vessel diseases with 7T MRI : Rationale and design of the " ZOOM @ SVDs " study. Cerebral Circulation - Cognition and Behavior, 2(April), 100013. https://doi.org/10.1016/j.cccb.2021.100013

Buchanan, C. R., Bastin, M. E., Ritchie, S. J., Liewald, D. C., Madole, J. W., Tucker-Drob, E. M., Deary, I. J., & Cox, S. R. (2020). The effect of network thresholding and weighting on structural brain networks in the UK Biobank. NeuroImage, 211(January), 116443. https://doi.org/10.1016/j.neuroimage.2019.116443

Buchanan, C. R., Pernet, C. R., Gorgolewski, K. J., Storkey, A. J., & Bastin, M. E. (2014). Test-retest reliability of structural brain networks from diffusion MRI. NeuroImage, 86, 231–243. https://doi.org/10.1016/j.neuroimage.2013.09.054

Bullmore, E., & Sporns, O. (2009). Complex brain networks : graph theoretical analysis of structural and functional systems. 10(February). https://doi.org/10.1038/nrn2575

Cetin-Karayumak, S., Di Biase, M. A., Chunga, N., Reid, B., Somes, N., Lyall, A. E., Kelly, S., Solgun, B., Pasternak, O., Vangel, M., Pearlson, G., Tamminga, C., Sweeney, J. A., Clementz, B., Schretlen, D., Viher, P. V., Stegmayer, K., Walther, S., Lee, J., … Kubicki, M. (2020). White matter abnormalities across the lifespan of schizophrenia: a harmonized multi-site diffusion MRI study. Molecular Psychiatry, 25(12), 3208–3219. https://doi.org/10.1038/s41380-019-0509-y

Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., & Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. NeuroImage, 184, 180–200. https://doi.org/10.1016/j.

neuroimage.2018.08.073

Craggs, L. J. L., Yamamoto, Y., Ihara, M., Fenwick, R., Burke, M., Oakley, A. E., Roeber, S., Kretzschmar, H., Neuropathology, R. N. K., & Neurobiology, A. (2014). White matter pathology and disconnection in the frontal lobe in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). 44, 591–602. https://doi.org/10.1111/nan.12073

de Brito Robalo, B. M., Jan Biessels, G., Chen, C., Dewenter, A., Duering, M., Hilal, S., Koek, H. L., Kopczak, A., Yin Ka Lam, B., Leemans, A., Mok, V., Onkenhout, L. P., van den Brink, H., & de Luca, A. (2021). Diffusion MRI harmonization enables joint-analysis of multicentre data of patients with cerebral small vessel disease. NeuroImage: Clinical, 32(October), 102886. https://doi.org/10.1016/j.nicl.2021.102886

de Reus, M. A., & van den Heuvel, M. P. (2013a). Estimating false positives and negatives in brain networks. NeuroImage, 70, 402–409. https://doi.org/10.1016/j.neuroimage.2012.12.066

de Reus, M. A., & van den Heuvel, M. P. (2013b). The parcellation-based connectome: Limitations and extensions. NeuroImage, 80, 397–404. https://doi.org/10.1016/j.neuroimage.2013.03.053

Di Donato, I., Bianchi, S., De Stefano, N., Dichgans, M., Dotti, M. T., Duering, M., Jouvent, E., Korczyn, A. D., Lesnik-Oberstein, S. A. J., Malandrini, A., Markus, H. S., Pantoni, L., Penco, S., Rufa, A., Sinanović, O., Stojanov, D., & Federico, A. (2017). Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL) as a model of small vessel disease: Update on clinical, diagnostic, and management aspects. BMC Medicine, 15(1), 1–12. https://doi.org/10.1186/s12916-017-0778-8

Du, J., Zhu, H., Zhou, J., Lu, P., Qiu, Y., Yu, L., Cao, W., Zhi, N., Yang, J., Xu, Q., Sun, J., & Zhou, Y. (2020). Structural Brain Network Disruption at Preclinical Stage of Cognitive Impairment Due to Cerebral Small Vessel Disease. Neuroscience, 449, 99–115. https://doi.org/10.1016/j.neuroscience.2020.08.037

Fazekas, F., Chawluk, J. B., & Alavi, A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. American Journal of Neuroradiology, 8(3), 421–426.

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. NeuroImage, 161(July), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Frey, B. M., Petersen, M., Schlemm, E., Mayer, C., Hanning, U., Engelke, K., Fiehler, J., Borof, K., Jagodzinski, A., Gerloff, C., Thomalla, G., & Cheng, B. (2021). White matter integrity and structural brain network topology in cerebral small vessel disease: The Hamburg city health study. Human Brain Mapping, 42(5), 1406–1415. https://doi.org/10.1002/hbm.25301

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Van Wedeen, J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. PLoS Biology, 6(7), 1479–1493. https://doi.org/10.1371/journal.pbio.0060159

Heinen, R., Vlegels, N., de Bresser, J., Leemans, A., Biessels, G. J., & Reijmer, Y. D. (2018). The cumulative effect of small vessel disease lesions is reflected in structural brain networks of memory clinic patients. NeuroImage: Clinical, 19(May), 963–969. https://doi.org/10.1016/j.nicl.2018.06.025

Jeurissen, B., Leemans, A., Jones, D. K., Tournier, J. D., & Sijbers, J. (2011). Probabilistic fiber

tracking using the residual bootstrap with constrained spherical deconvolution. Human Brain Mapping, 32(3), 461–479. https://doi.org/10.1002/hbm.21032

Kurokawa, R., Kamiya, K., Koike, S., Nakaya, M., Uematsu, A., Tanaka, S. C., Kamagata, K., Okada, N., Morita, K., Kasai, K., & Abe, O. (2021). Cross-scanner reproducibility and harmonization of a diffusion MRI structural brain network: A traveling subject study of multi-b acquisition. NeuroImage, 245(October), 118675. https://doi.org/10.1016/j.neuroimage.2021.118675

Lam, B. Y. K., Leung, K. T., Yiu, B., Zhao, L., Biesbroek, J. M., Au, L., Tang, Y., Wang, K., Fan, Y., Fu, J. H., Xu, Q., Song, H., Tian, X., Chu, W. C. W., Abrigo, J., Shi, L., Ko, H., Lau, A., Duering, M., … Mok, V. C. T. (2019). Peak width of skeletonized mean diffusivity and its association with age-related cognitive alterations and vascular risk factors. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, 11, 721–729. https://doi.org/10.1016/j.dadm.2019.09.003

Lawrence, A. J., Zeestraten, E. A., Benjamin, P., Lambert, C. P., Morris, R. G., Barrick, T. R., & Markus, H. S. (2018). Longitudinal decline in structural networks predicts dementia in cerebral small vessel disease. Neurology, 90(21), e1898–e1910. https://doi.org/10.1212/WNL.0000000000005551

Leemans, A., Jeruissen, B., Sijbers, J., & Jones, D. K. (2009). ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. Proceedings of the International Society for Magnetic Resonance in Medicine, 245(2), 3537. https://doi.org/10.1093/occmed/kqr069

Leemans, A., & Jones, D. K. (2009). The B-matrix must be rotated when correcting for subject motion in DTI data. Magnetic Resonance in Medicine, 61(6), 1336–1349. https://doi.org/10.1002/mrm.21890

McColgan, P., Blom, T., Rees, G., Seunarine, K., Gregory, S., Johnson, E., Durr, A., Roos, R., Scahill, R., Clark, C., Tabrizi, S., & Razi, A. (2018). Stability and sensitivity of structural connectomes: effect of thresholding and filtering and demonstration in neurodegeneration. BioRxiv, 416826. https://doi.org/10.1101/416826

Messaritaki, E., Dimitriadis, S. I., & Jones, D. K. (2019). Optimization of graph construction can significantly increase the power of structural brain network studies. NeuroImage, 199, 495–511. https://doi.org/10.1016/j.neuroimage.2019.05.052

Mirzaalian, H., De Pierrefeu, A., Savadjiev, P., Pasternak, O., Bouix, S., Kubicki, M., Westin, C. F., Shenton, M. E., & Rathi, Y. (2015). Harmonizing diffusion MRI data across multiple sites and scanners. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9349, 12–19. https://doi.org/10.1007/978-3-319-24553-9_2

Petersen, M., Frey, B. M., Schlemm, E., Mayer, C., Hanning, U., Engelke, K., Fiehler, J., Borof, K., Jagodzinski, A., Gerloff, C., Thomalla, G., & Cheng, B. (2020). Network Localisation of White Matter Damage in Cerebral Small Vessel Disease. Scientific Reports, 10(1), 1–9. https://doi.org/10.1038/s41598-020-66013-w

Pievani, M., Filippini, N., Van Den Heuvel, M. P., Cappa, S. F., & Frisoni, G. B. (2014). Brain connectivity in neurodegenerative diseases - From phenotype to proteinopathy. Nature Reviews Neurology, 10(11), 620–633. https://doi.org/10.1038/nrneurol.2014.178

Reijmer, Y. D., Fotiadis, P., Martinez-Ramirez, S., Salat, D. H., Schultz, A., Shoamanesh, A.,

Ayres, A. M., Vashkevich, A., Rosas, D., Schwab, K., Leemans, A., Biessels, G. J., Rosand, J., Johnson, K. A., Viswanathan, A., Gurol, M. E., & Greenberg, S. M. (2015). Structural network alterations and neurological dysfunction in cerebral amyloid angiopathy. Brain, 138(1), 179–188. https://doi.org/10.1093/brain/awu316

Reijmer, Y. D., Fotiadis, P., Piantoni, G., Boulouis, G., Kelly, K. E., Gurol, M. E., Leemans, A., O'Sullivan, M. J., Greenberg, S. M., & Viswanathan, A. (2016). Small vessel disease and cognitive impairment: The relevance of central network connections. Human Brain Mapping, 37(7), 2446–2454. https://doi.org/10.1002/hbm.23186

Reijmer, Y. D., Leemans, A., Brundel, M., Kappelle, L. J., & Biessels, G. J. (2013). Disruption of the cerebral white matter network is related to slowing of information processing speed in patients with type 2 diabetes. Diabetes, 62(6), 2112–2115. https://doi.org/10.2337/db12-1644

Robalo, B. M. D. B., Vlegels, N., Meier, J., Leemans, A., & Biessels, G. J. (2020). Effect of Fixed-Density Thresholding on Structural Brain Networks : A Demonstration in Cerebral Small Vessel Disease. https://doi.org/10.1089/brain.2019.0686

Roberts, J. A., Perry, A., Roberts, G., Mitchell, P. B., & Breakspear, M. (2017). Consistency-based thresholding of the human connectome. NeuroImage, 145(March 2016), 118–129. https://doi.org/10.1016/j.neuroimage.2016.09.053

Schilling, K. G., Tax, C. M. W., Rheault, F., Hansen, C., Yang, Q., Yeh, F. C., Cai, L., Anderson, A. W., & Landman, B. A. (2021). Fiber tractography bundle segmentation depends on scanner effects, vendor effects, acquisition resolution, diffusion sampling scheme, diffusion sensitization, and bundle segmentation workflow. NeuroImage, 242(July), 118451. https://doi.org/10.1016/j.neuroimage.2021.118451

Smith, E. E., Biessels, G. J., De Guio, F., de Leeuw, F. E., Duchesne, S., Düring, M., Frayne, R., Ikram, M. A., Jouvent, E., MacIntosh, B. J., Thrippleton, M. J., Vernooij, M. W., Adams, H., Backes, W. H., Ballerini, L., Black, S. E., Chen, C., Corriveau, R., DeCarli, C., … Wardlaw, J. M. (2019). Harmonizing brain magnetic resonance imaging methods for vascular contributions to neurodegeneration. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, 11, 191–204. https://doi.org/10.1016/j.dadm.2019.01.002

Smith RE, Tournier JD, Calamante F, Connelly A. SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. Neuroimage. 2015 Oct 1;119:338-51. doi: 10.1016/j.neuroimage.2015.06.092. Epub 2015 Jul 8. PMID: 26163802.

Sotiropoulos, S. N., Jbabdi, S., Xu, J., Andersson, J. L., Moeller, S., Auerbach, E. J., Glasser, M. F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D. A., Yacoub, E., Lenglet, C., Van Essen, D. C., Ugurbil, K., & Behrens, T. E. J. (2013). Advances in diffusion MRI acquisition and processing in the Human Connectome Project. NeuroImage, 80, 125–143. https://doi.org/10.1016/j.neuroimage.2013.05.057

Tax, C. M. W., Otte, W. M., Viergever, M. A., Dijkhuizen, R. M., & Leemans, A. (2015). REKINDLE: Robust Extraction of Kurtosis INDices with Linear Estimation. Magnetic Resonance in Medicine, 73(2), 794–808. https://doi.org/10.1002/mrm.25165

Tournier, J. D., Yeh, C. H., Calamante, F., Cho, K. H., Connelly, A., & Lin, C. P. (2008). Resolving crossing fibres using constrained spherical deconvolution: Validation using diffusion-weighted imaging phantom data. NeuroImage, 42(2), 617–625. https://doi.org/10.1016/j.neuroimage.2008.05.002

Tuladhar, A. M., Lawrence, A., Norris, D. G., Barrick, T. R., Markus, H. S., & de Leeuw, F. E. (2017). Disruption of rich club organisation in cerebral small vessel disease. Human Brain Mapping, 38(4), 1751–1766. https://doi.org/10.1002/hbm.23479

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage, 15(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. Trends in Cognitive Sciences, 17(12), 683–696. https://doi.org/10.1016/j.tics.2013.09.012

van Wijk, B. C. M., Stam, C. J., & Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. PLoS ONE, 5(10). https://doi.org/10.1371/journal.pone.0013701

F. Váša, E.T. Bullmore, A.X. Patel Probabilistic thresholding of functional connectomes: application to schizophrenia Neuroimage, 172 (December 2017) (2018), pp. 326-340, 10.1016/j.neuroimage.2017.12.043

Veraart, J., Sijbers, J., Sunaert, S., Leemans, A., & Jeurissen, B. (2013). Weighted linear least squares estimation of diffusion MRI parameters: Strengths, limitations, and pitfalls. NeuroImage, 81, 335–346. https://doi.org/10.1016/j.neuroimage.2013.05.028

Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., Duncan, J. S., Richardson, M. P., & Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. NeuroImage, 51(4), 1384–1394. https://doi.org/10.1016/j.neuroimage.2010.03.046

Vos, S. B., Tax, C. M. W., Luijten, P. R., Ourselin, S., Leemans, A., & Froeling, M. (2017). The importance of correcting for signal drift in diffusion MRI. Magnetic Resonance in Medicine, 77(1), 285–299. https://doi.org/10.1002/mrm.26124

Yeh, C. H., Jones, D. K., Liang, X., Descoteaux, M., & Connelly, A. (2021). Mapping Structural Connectivity Using Diffusion MRI: Challenges and Opportunities. Journal of Magnetic Resonance Imaging, 53(6), 1666–1682. https://doi.org/10.1002/jmri.27188

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. NeuroImage, 53(4), 1197–1207. https://doi.org/10.1016/j.neuroimage.2010.06.041

Chapter 6

Summary and General Discussion

SUMMARY

The aim of the work described in this thesis was to advance the implementation of diffusion MRI in SVD research by exploring methods that improve its technical validity. The two key objectives were 1) to improve the validity of multicentre dMRI, by removing variability related to acquisition hardware in dMRI across sites with harmonization of the raw diffusion signal and 2) to improve the validity of dMRI-based network analyses, in the presence of pathology, by removing false positive connections with thresholding.

We addressed our first key objective in Chapters 2 and 5, where we showed proof of concept of effective harmonization with data from five cohorts of patients with SVD and controls, using rotation invariant spherical harmonic (RISH) features. In **chapter 2** we showed that the RISH method removes acquisition-related differences in scalar dMRI metrics between matched controls of different sites, while preserving disease-related effect sizes (i.e., differences between patients and controls and associations between dMRI metrics and markers of SVD). Consequently, the harmonized data could be pooled to increase sample size and infer associations between dMRI metrics and markers of SVD with improved power. In **chapter 5** we applied RISH harmonization beyond the scalar metrics explored in Chapter 2, by investigating if it also improves cross-site consistency of brain networks. We demonstrated that harmonization helps to achieve more similar network architectures across sites. Furthermore, harmonization facilitated data pooling to infer patterns of network injury with improved sensitivity as compared to single centre datasets.

We addressed our second key objective in chapters 3, 4 and 5 where we examined whether network thresholding increases consistency of brain networks in studies with cross-sectional design, longitudinal design and in multicentre analysis. In **chapter 3** we proposed fixed-density thresholding as method to control for differences in network density (i.e., number of detected connections) when comparing networks of patients and controls. We showed that thresholding networks to a fixed density across all subjects preserves the original network architecture over a large range of thresholds and maintains the sensitivity to detected group-differences between patients with SVD and controls in global and local metrics. In this manner, networks can be compared across groups without the risk of measures being affected by density bias. In **chapter 4,** we tackled the effect of false positive connections causing low reproducibility of brain networks in longitudinal studies. We showed that weight-based thresholding improves scan-rescan network consistency in healthy young subjects (with relatively short-rescan inter-

vals) but also in patients with SVD scanned over long time periods. Importantly, we proved that thresholding preserves sensitivity to detect statistical group-differences between patients with low and high SVD burden. Finally, in **chapter 5** we demonstrated that thresholding, in combination with RISH harmonization, helps to generate more consistent networks in multicentre data. In this analysis we showed that thresholding complements harmonization by reducing the number of false positives in the network, ultimately improving precision to detect patterns of network injury in multicentre data of patients SVD.

Altogether, the work presented in this thesis shows the benefits of dMRI harmonization by making data more comparable across centres and enabling data pooling to increase sample size. Our work also shows that advantages of network thresholding by improving precision of network analysis, paving the way for robust large scale dMRI network analysis in SVD.

DISCUSSION

Dealing with scanner and acquisition-related differences in diffusion MRI

The future use of dMRI as a diagnostic tool in studies of neurological and psychiatric disorders requires reliability and stability of measures across scanners, and acquisition parameters (Teipel et al., 2011). The need for data harmonization has increased even more with the availability of large multicentre datasets such as the Human Connectome Project (HCP) (Essen et al., 2013) or the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2015). To take full advantage of such datasets where new imaging markers can be discovered and validated, harmonization strategies that deal with unwanted variability need to be included in the pipeline, prior to data pooling and joint analysis.

An important choice we faced when implementing harmonization was whether to harmonize the raw data or individual diffusion metrics. Given that SVD is a condition associated with changes in multiple dMRI metrics derived from different models and analysis pipelines (e.g., FA, MD, van Norden et al., 2012; PSMD, Baykara et al., 2016; tractography and connectivity metrics, Petersen et al., 2020), it was more advantageous to harmonize the raw data. The major disadvantage of the alternative approach, harmonizing the individual diffusion metrics, is that the choice of what kind of analysis and inferences we could perform afterwards would be very limited. Furthermore, harmonizing individual metrics could have implications for multivariate analyses, since it is not guaranteed that well known relations between those metrics (e.g., high FA in combination with low MD) are preserved. We chose the RISH method, which harmonizes the raw data in a model independent manner, therefore allowing us to calculate any diffusion metric in subsequent analyses (e.g., TBSS and voxel-wise comparison in Chapter 2 and network analysis in Chapter 5). We recognize that other harmonization methods that work on the raw data also exist. For example, the method of moments, which harmonizes a target site to a reference site by matching their spherical mean and spherical variance to correct for unwanted variability. Another recently method is sparse dictionary learning which uses machine learning approaches for harmonization (St-Jean et al., 2019). We chose to implement the RISH since it performed best in a recent challenge on dMRI harmonization (Lipeng et al). The best harmonization method for a particular study depends on the study design, the research question, the available data and the known sources of variability as discussed below. In Figure 1, we summarize questions that we faced when choosing which method to implement, and we recommend following this chart when choosing a harmonization method.

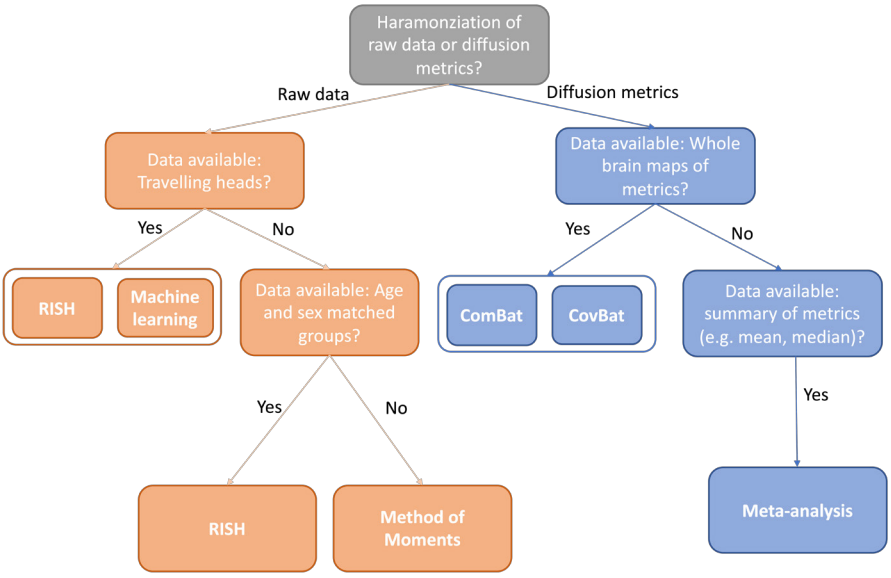


Figure 1: Illustration of potential questions when implementing a harmonization pipeline. The first question is whether harmonization is performed on the raw data (left side - orange) or on individual diffusion MRI metrics (right side). Left: For harmonization of raw data, the most suitable method depends on what kind of the training data is available. Do we have access to scans of the same subjects scanned on all sites ("travelling heads")? If yes, machine learning approaches such as Sparse Dictionary Learning (St-Jean et al., 2019) and a recent iteration of the RISH method (Cetin Karayumak et al., 2019) can be applied. If travelling subjects are not available, can we define age- and sex-matched groups across all sites? If yes, the RISH method is again suitable and if no, the Method of Moments (Huynh et al., 2019) is the alternative. Right: For harmonization of diffusion metrics, do we have access to whole brain maps of each subject? If yes, approaches that use regression of covariates such as ComBat (Fortin et al., 2017) are recommended. Otherwise, if only summary of statistics available, we are left with Meta-analysis approaches (Teipel et al., 2012).

When implementing the RISH approach, a challenge we had to consider was how to accurately match subjects across cohorts to train the harmonization algorithm. In the training phase, a key assumption is that statistical differences in dMRI between matched groups only occur due to scanner and acquisition-related differences. Consequently, if matching is optimal in this respect, the dMRI signal can be harmonized between two sites by calculating scaling factors between the average of RISH features of matched subjects included. In the studies presented in Chapters 2 and 5, we had access to scans of elderly controls, which were established as training subjects and matched across sites not only in terms of age and sex. Yet, other factors that should ideally be considered are ethnicity, handedness, education level, but also other potential factors that influence the diffusion signal such as the degree of brain lesions.

During our selection process we ensured that training controls only included subjects with minimal degree of WMH and no major artefacts. In an ideal scenario, all harmonization studies should have "travelling subjects" that are scanned on all sites, thus avoiding the hurdles of group

matching. We showed that when traveling subjects are not available, recruiting a group of 15 matched subjects at each site is sufficient to effectively remove acquisition-related differences. Future studies implementing the RISH method on clinical cohorts should also consider that anatomical alterations from their relevant condition (e.g., tumour or stroke), particularly in the training data, could severely affect the performance of harmonization. In those cases, **we recommend to a least ensure that the location of anatomical alterations is the same across all sites**.

After computing harmonized scans, the question that followed was how to establish if harmonization was successful. For this purpose, we examined the preservation of disease-related effects alongside removal of scanner-related differences. We demonstrated that statistical differences in dMRI metrics between the training controls are removed, and that the expected variability between patients with SVD and controls as well as associations between dMRI metrics and SVD markers are preserved. This was true not only for metrics derived from the diffusion tensor model (FA, MD, PSMD in Chapter 2), but also for metrics obtained with more sophisticated analysis such as fibre tractography and network analysis (Chapter 5). We thus proved that RISH harmonization, which is applied at the beginning of the processing pipeline, generates more comparable dMRI metrics between sites, regardless of which approach is used to calculate those metrics. Consequently, data could be pooled at different stages of processing to increase sample size and confidence to test distinct effects of interest (e.g., associations between dMRI and WMH in Chapter 2 or patterns of network injury in Chapter 5). Similar validations of the RISH method have been shown in controls and in a large population of patients with schizophrenia, where statistical differences due to scanner were removed in matched groups, while preserving biological variation due to age and sex (Cetin-Karayumak et al., 2020). When implementing RISH harmonization with patient populations, **we recommend verifying whether well-established disease effects** are preserved, as part of the quality assessment of the processing pipeline. If traveling subjects were available, this would have provided a ground truth to directly compare RISH features or resulting dMRI metrics between sites on a subject-by-subject basis, thus simplifying the validation process.

Despite the benefits of RISH harmonization demonstrated in this work, some challenges remain before dMRI can be fully implemented in clinical SVD research. Ideally, harmonization should not require training data and should be performed directly on patients since many studies do not include healthy controls alongside their patient population. Currently, it is only possible to harmonize data with relatively similar spatial resolutions, b-value and number of diffusion gradient directions(Tax et al., 2019). Beyond harmonization, other steps can also be taken to facilitate multicentre data sharing and minimize scanner and acquisition-related differences. Ideally, in multicentre studies, all scans should be acquired with the same scanner hardware acquisition parameters and similar calibration phantoms from the beginning of the study. However, that is both a financial and logistical challenge. Nonetheless, at the very least, scanner vendors and researchers have the responsibility of being transparent with regards to reconstruction algorithms, acquisition protocols and processing steps.

Improving consistency of brain network analysis in the presence of disease

The study of brain networks has the potential to advance the understanding of many psychiatric and neurological disorders. However, even within single research fields such as SVD, there are large differences in defining network architectures, and disease effects. Apart from removing scanner-related differences in the raw data as proposed above, during network analysis we deal with other types of unwanted variability between datasets, whether we are comparing networks in cross-sectional studies, longitudinal studies or in multicentre analysis. In all scenarios the main challenge is that brain networks contain many false positive connections.

As shown in Chapter 3, in single-centre cross-sectional analysis, the presence of false positives translates into significant differences in network density across subjects, which can bias group comparisons of network metrics such as global efficiency. This is especially problematic in datasets of elderly subjects where fibre tractography is suboptimal because of brain lesions, lower scan quality and misregistration, resulting in many false positives. In chapter 3 we applied fixed-density thresholding to remove false positives and generate networks with fixed densities, so that eventual differences in global network measures across groups appear solely through structural changes rather than density differences. We noticed that the disadvantage of this approach is the manipulation of the empirical network by removing connections since it could be possible that differences in overall density between subjects is caused by the experimental condition rather than a mere confounder. For this reason, **we recommend checking whether the original network architecture (e.g., hub location, hub-hierarchy and connection weights) is maintained over different threshold levels**. Moreover, group-differences that reflect real disease effects should be observed over a large range of thresholds and not only at a specific threshold.

As an alternative to fixed-density thresholding, density bias could also be dealt with by normalizing network measures to random graphs (van Wijk et al., 2010). However, that also comes with its challenges since normalization introduces different bias for local and global network measures. This means that a uniform normalization step cannot be applied to all metrics at once. Until ground-truth brain networks have been developed both solutions will have their own limitations.

Regarding longitudinal studies, the presence of false positives contributes for low scan-rescan network architecture reproducibility when subjects are scanned over multiple time points, especially when rescan intervals span months or years. The question we faced at this stage was how can we ensure that subtle changes over time caused by real disease effects and not confounded by low scan-rescan reproducibility? In the research presented in Chapter 4, we demonstrated that thresholding reduces the number of false positives in the network and generates more consistent network architectures over time (e.g., hub location and edges detected). Thresholded networks had a more consistent scan-rescan structure over which disease effects over time could be evaluated. However, we must consider that improving consistency of network architecture is only one side of the coin since thresholding could potentially remove genuine connections that

are relevant for assessing disease effects. As we demonstrated, one way to determine whether disease effects are preserved is to **examine the interindividual variation of network metrics and associations with external markers (e.g., burden of WMH) over different threshold levels.** We tested the performance of two weight-based thresholding methods (fixed-density and absolute thresholding), but we are aware that more thresholding methods have been proposed (e.g., consensus and consistency thresholding (Betzel et al., 2018; Roberts et al., 2017). These methods were developed using datasets of young subjects, but it is also important to test their validity on scans of patients, especially when we the longitudinal analysis is focused on group-level networks rather than individual networks as described in Chapter 4. Other recently developed thresholding approaches directly filter individual streamlines from the tractography data, which could also improve scan-rescan consistency of network architecture in disease (Smith et al., 2015). Apart from thresholding, in longitudinal studies scan-rescan reproducibility can also be improved by using the same tractography algorithms, parcellation schemes and network weights (de Reus & van den Heuvel, 2013b).

In multicentre network analysis, the presence of false positives reduces consistency of networks across centres and restricts data pooling. With the study described in Chapter 5, we showed that combining harmonization and thresholding generates networks with more consistent architectures across sites by removing low probability tracts. Even if no ground-truth brain network has been established, it is reasonable to assume that connections that represent true underlying white matter pathways will be consistently reconstructed across experiments. Thresholding helped to achieve this robust structure containing only the most consistent connections, over which disease effects could eventually be assessed. When working with multicentre data, even though thresholding alone improves the consistency of which connections are reconstructed, **we recommend combining it with harmonization so that the connections weights of those connections also become more similar across sites**. Only then can the data be pooled to increase sample size. We demonstrated that combining thresholding, harmonization and multicentre data pooling significantly improves the sensitivity and precision to detect patterns of network injury in patients with SVD. For network findings to be useful in clinical studies, they need be validated in large datasets, particularly if effect sizes are expected to be small.

CONCLUSIONS

The work presented in this thesis explored methods to improve the technical validity of diffusion MRI and advance its implementation in SVD research. We demonstrated that harmonization with the RISH method removed scanner and acquisition-related differences between diffusion scans acquired on different centres. The harmonization pipeline, as presented in this thesis, can be implemented in future SVD studies with multicentre data, to increase sample size and sensitivity to disease affects. Importantly, future studies should be aware that subject matching will be a challenging step, especially if the training controls are elderly subjects or contain brain lesions. In that case we recommend using scans where brain lesions are comparable across

sites, both in terms of location and severity. We also recommend verifying whether well-known biological effects are preserved after harmonization, alongside removal of cross-site differences in the training group.

We also showed that network thresholding improves consistency of brain networks derived from dMRI, while preserving disease effects in cross-sectional and longitudinal analyses. For studies with a similar design, we propose applying the thresholding approaches discussed in this work to obtain more consistent network architectures and increase precision of network analysis. We recommend only applying threshold levels where the original network architecture and the interindividual variation of subjects are preserved. We also demonstrated benefits of thresholding in multicentre analysis, thresholding in generating more comparable networks across centres and in increasing precision to detect disease effects.

I hope that the ideas presented in this thesis will guide researchers setting up dMRI studies in clinical settings, with a vision to validate their findings in large datasets.

REFERENCES

Baykara, E., Gesierich, B., Adam, R., Tuladhar, A. M., Biesbroek, J. M., Koek, H. L., Ropele, S., Jouvent, E., Chabriat, H., Ertl-Wagner, B., Ewers, M., Schmidt, R., de Leeuw, F. E., Biessels, G. J., Dichgans, M., & Duering, M. (2016). A Novel Imaging Marker for Small Vessel Disease Based on Skeletonization of White Matter Tracts and Diffusion Histograms. Annals of Neurology, 80(4), 581–592. https://doi.org/10.1002/ana.24758

Betzel, R. F., Griffa, A., Hagmann, P., & Mišić, B. (2018). Distance-dependent consensus thresholds for generating group-representative structural brain networks. Network Neuroscience, 1–22. https://doi.org/10.1162/netn_a_00075

Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., & Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. NeuroImage, 184, 180–200. https://doi.org/10.1016/j.neuroimage.2018.08.073

Cetin-Karayumak, S., di Biase, M. A., Chunga, N., Reid, B., Somes, N., Lyall, A. E., Kelly, S., Solgun, B., Pasternak, O., Vangel, M., Pearlson, G., Tamminga, C., Sweeney, J. A., Clementz, B., Schretlen, D., Viher, P. V., Stegmayer, K., Walther, S., Lee, J., … Kubicki, M. (2020). White matter abnormalities across the lifespan of schizophrenia: a harmonized multi-site diffusion MRI study. Molecular Psychiatry, 25(12), 3208–3219. https://doi.org/10.1038/s41380-019-0509-y

de Reus, M. A., & van den Heuvel, M. P. (2013). The parcellation-based connectome: Limitations and extensions. NeuroImage, 80, 397–404. https://doi.org/10.1016/j.neuroimage.2013.03.053

Essen, D. C. van, Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & Consortium, W. H. C. P. (2013). NeuroImage The WU-Minn Human Connectome Project : An overview. 80, 62–79.

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. NeuroImage, 161(July), 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Huynh, K. M., Chen, G., Wu, Y., Shen, D., & Yap, P. T. (2019). Multi-Site Harmonization of Diffusion MRI Data via Method of Moments. IEEE Transactions on Medical Imaging, 38(7), 1599–1609. https://doi.org/10.1109/TMI.2019.2895020

Jack, C. R., Barnes, J., Bernstein, M. A., Borowski, B. J., Brewer, J., Clegg, S., Dale, A. M., Carmichael, O., Ching, C., DeCarli, C., Desikan, R. S., Fennema-Notestine, C., Fjell, A. M., Fletcher, E., Fox, N. C., Gunter, J., Gutman, B. A., Holland, D., Hua, X., … Weiner, M. (2015). Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. Alzheimer's and Dementia, 11(7), 740–756. https://doi.org/10.1016/j.jalz.2015.05.002

Petersen, M., Frey, B. M., Schlemm, E., Mayer, C., Hanning, U., Engelke, K., Fiehler, J., Borof, K., Jagodzinski, A., Gerloff, C., Thomalla, G., & Cheng, B. (2020). Network Localisation of White Matter Damage in Cerebral Small Vessel Disease. Scientific Reports, 10(1), 1–9. https://doi.org/10.1038/s41598-020-66013-w

Roberts, J. A., Perry, A., Roberts, G., Mitchell, P. B., & Breakspear, M. (2017). Consistency-based thresholding of the human connectome. NeuroImage, 145(September 2016), 118–129. https://doi.org/10.1016/j.neuroimage.2016.09.053

Smith, R. E., Tournier, J. D., Calamante, F., & Connelly, A. (2015). SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. NeuroImage, 119, 338–351. https://doi.org/10.1016/j.neuroimage.2015.06.092

St-Jean, S., Viergever, M. A., & Leemans, A. (2019). Harmonization of diffusion MRI datasets with adaptive dictionary learning. http://arxiv.org/abs/1910.00272

Tax, C. M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., John Evans, C., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., Ghosh, A., Tanno, R., Alexander, D. C., Zappalà, S., Charron, C., Kusmia, S., Linden, D. E., Jones, D. K., & Veraart, J. (2019). Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. NeuroImage, 195(February), 285–299. https://doi.org/10.1016/j.neuroimage.2019.01.077

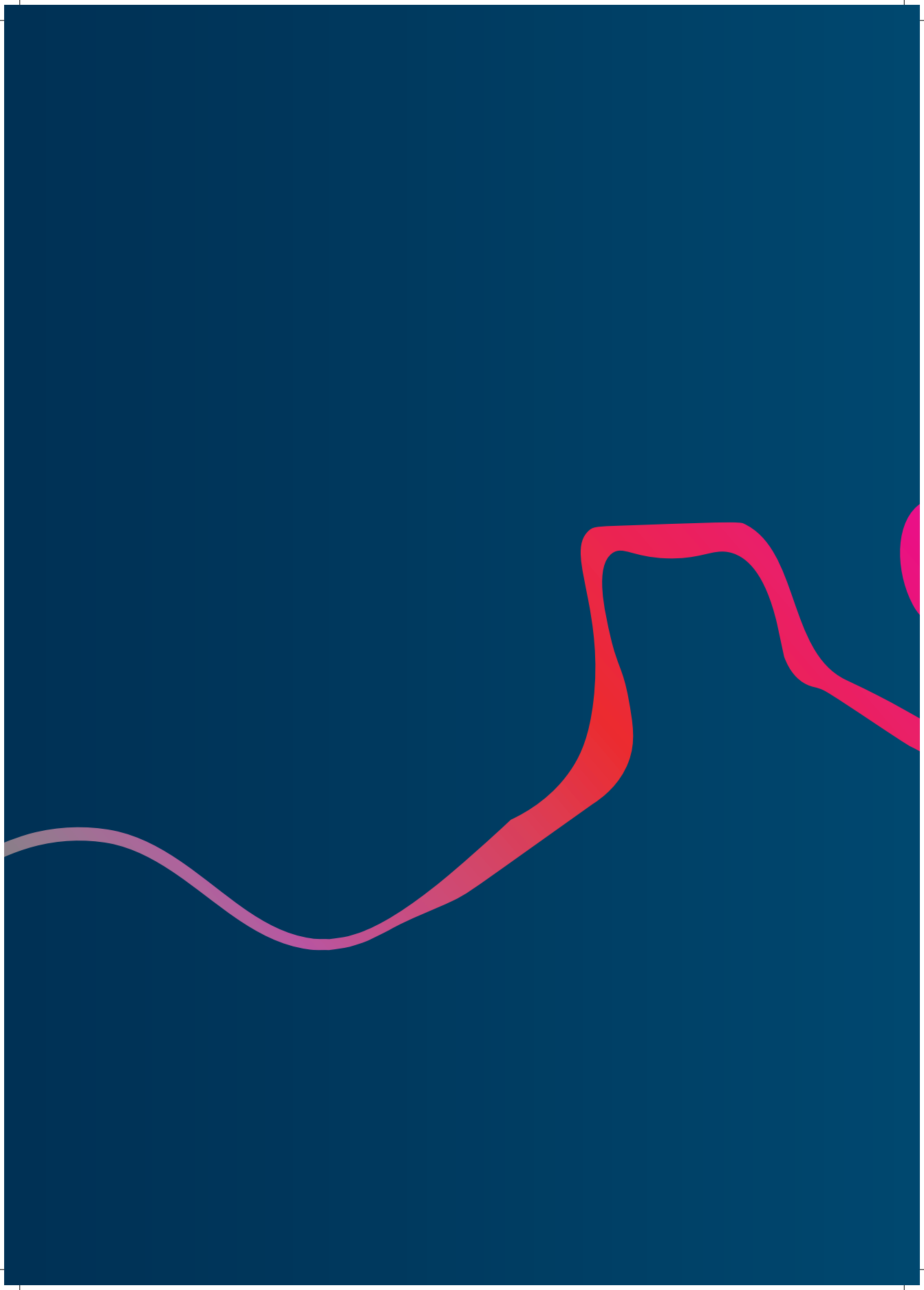
Teipel, S. J., Reuter, S., Stieltjes, B., Acosta-Cabronero, J., Ernemann, U., Fellgiebel, A., Filippi, M., Frisoni, G., Hentschel, F., Jessen, F., Klöppel, S., Meindl, T., Pouwels, P. J. W., Hauenstein, K. H., & Hampel, H. (2011). Multicenter stability of diffusion tensor imaging measures: A European clinical and physical phantom study. Psychiatry Research - Neuroimaging, 194(3), 363–371. https://doi.org/10.1016/j.pscychresns.2011.05.012

Teipel, S. J., Wegrzyn, M., Meindl, T., Frisoni, G., Bokde, A. L. W., Fellgiebel, A., Filippi, M., Hampel, H., Klöppel, S., Hauenstein, K., & Ewers, M. (2012). Anatomical MRI and DTI in the diagnosis of Alzheimer's disease: A european multicenter study. Journal of Alzheimer's Disease, 31(SUPPL. 3). https://doi.org/10.3233/JAD-2012-112118

Tuladhar, A. M., van Dijk, E., Zwiers, M. P., van Norden, A. G. W., de Laat, K. F., Shumskaya, E., Norris, D. G., & de Leeuw, F. E. (2016). Structural network connectivity and cognition in cerebral small vessel disease. Human Brain Mapping, 37(1), 300–310. https://doi.org/10.1002/hbm.23032

van Norden, A. G. W., de Laat, K. F., van Dijk, E. J., van Uden, I. W. M., van Oudheusden, L. J. B., Gons, R. A. R., Norris, D. G., Zwiers, M. P., & de Leeuw, F. E. (2012). Diffusion tensor imaging and cognition in cerebral small vessel disease. The RUN DMC study. Biochimica et Biophysica Acta - Molecular Basis of Disease, 1822(3), 401–407. https://doi.org/10.1016/j.bbadis.2011.04.008

van Wijk, B. C. M., Stam, C. J., & Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. PLoS ONE, 5(10). https://doi.org/10.1371/journal.pone.0013701



Appendices

Nederlandse samenvatting
Abbreviations
Publications
Acknowledgements (Dankwoord)
About the author

NEDERLANDSE SAMENVATTING

Diffusie-MRI (dMRI) is een veelbelovende techniek om hersenschade te detecteren bij patienten met small vessel disease (SVD). Zo is met behulp van dMRI aangetoond dat bepaalde verbindingen in het hersennetwerk van patiënten met SVD zijn beschadigd en dat deze schade samenhangt met cognitieve problemen. Helaas kent de techniek momenteel nog een aantal technische beperkingen die grootschalig onderzoek en implementatie in de weg staan. Het is bijvoorbeeld niet goed mogelijk om dMRI gegevens afkomstig van verschillende centra samen te voegen, omdat elk type scanner een ander effect heeft op het diffusie signaal. Daarnaast bevatten op dMRI-gebaseerde hersennetwerken relatief veel vals-positieve verbindingen, waardoor ze niet goed te reproduceren zijn. Dit maakt de vergelijking van netwerken over tijd en tussen groepen minder nauwkeurig.

Het onderzoek in dit proefschrift richt zich op het verbeteren van de toepasbaarheid van dMRI in SVD door de bovenstaande technische beperkingen van deze techniek te verbeteren. De twee hoofddoelstellingen waren:

1) Het minimaliseren van centra-gerelateerde variatie in het dMRI signaal door een methode toe te passen die het onbewerkt diffusiesignaal 'harmoniseert'. 2) Het verbeteren van de reproduceerbaarheid van op dMRI gebaseerde hersennetwerken door het toepassen van verschillende 'filters'.

De eerste doelstelling hebben we onderzocht in **hoofdstuk 2 en 5**. Daarin hebben we de effectiviteit van de harmonisatie methode (RISH) laten zien in vijf verschillende groepen van patiënten met SVD en controles. In **hoofdstuk 2** laten we zien dat de RISH-methode inderdaad de centra-gerelateerde variatie in het dMRI signaal verwijdert, terwijl de ziekte-gerelateerde verschillen behouden blijven (d.w.z. verschillen tussen patiënten en controles en verschillen tussen patiënten met milde en ernstige SVD). Dit betekent dat vervolgstudies met deze methode grotere steekproeven kunnen onderzoeken, doordat dMRI beelden van verschillende centra bij elkaar gevoegd kunnen worden.

In **hoofdstuk 5** laten we zien dat de RISH-methode het ook mogelijk maakt om hersennetwerken samen te voegen die gebaseerd zijn op dMRI scans afkomstig van verschillende centra. We toonden aan dat schade aan het netwerk in patiënten met SVD eerder kan worden aangetoond in deze samengestelde steekproef dan op basis van gegevens afkomstig van één enkele locatie. De RISH-methode verbetert dus de gevoeligheid van dMRI om SVD-gerelateerde schade aan het hersennetwerk te detecteren.

De tweede doelstelling was het verbeteren van de reproduceerbaarheid van het hersennetwerk, door het verminderen van het aantal vals-positieve verbindingen. Dit hebben we onderzocht in **hoofdstuk 3, 4 en 5** door verschillende 'filters' toe te passen en het effect hiervan te bestuderen op de vergelijking tussen patiënten en controles (**hoofdstuk 3**), tussen 2 tijdspunten (**hoofdstuk**

4) en tussen patiënten en controles van scans die gemaakt zijn in meerdere centra (**hoofdstuk 5**).

In **hoofdstuk 3** hebben we een filter methode toegepast die ervoor zorgt dat elk netwerk uit een vast aantal 'kern' verbindingen bestaat. Met deze filter methode minimaliseer je het risico dat verschillen tussen groepen worden veroorzaakt door verschillen in de 'dichtheid' van het netwerk, oftewel het aantal verbindingen. We toonden aan dat ook als de dichtheid van het netwerk constant wordt gehouden, patiënten met SVD nog steeds gemiddeld een zwakker netwerk hebben dan controles.

In **hoofdstuk 4** hebben we een filter toegepast die het aantal vals-positieve verbindingen in het hersennetwerk vermindert door de meest zwakke verbindingen te verwijderen. We zagen na toepassing van deze filter inderdaad een verbetering van de reproduceerbaarheid van het hersennetwerk van gezonde jonge proefpersonen die op 2 tijdspunten waren gemeten (met een relatief kort tijdsinterval). Maar ook bij patiënten met SVD die over een langere perioden 2 keer waren gemeten waren de netwerken meer vergelijkbaar. Belangrijke hierbij is dat de gevoeligheid voor het detecteren van SVD-gerelateerde afwijkingen met deze methode werd behouden.

Ten slotte hebben we in **hoofdstuk 5** de filter methode uit hoofdstuk 4 gecombineerd met de RISH-harmonisatie methode om meer consistente hersennetwerken te genereren uit multicenter data. De resultaten lieten zien dat de combinatie van deze methoden het aantal vals-positieve verbindingen in het netwerk nog verder verminderd en meer consistente netwerken oplevert. Hierdoor is het mogelijk om nog nauwkeuriger afwijkingen aan het hersennetwerk in patiënten met SVD te detecteren.

Samenvattend toont het werk in dit proefschrift aan de harmonisatie- en filter-methoden de technische validiteit van dMRI in SVD onderzoek verbeterd. De methoden maken het mogelijk om dMRI gegevens uit verschillende centra samen te voegen om zo de steekproefomvang te vergroten. Daarnaast verbetert de filter methode de nauwkeurigheid van hersennetwerken die op basis van dMRI scans worden gereconstrueerd. Deze resultaten maken de weg vrij voor robuuste grootschalige dMRI-netwerkanalyse in SVD.

ABBREVIATIONS

AAL - Automated anatomical labelling

ADNI – Alzheimer's disease neuroimaging initiative

ANOVA – Analysis of variance

ANTs - Advanced Normalization Tools

BET – Brain extraction toolbox

CADASIL – Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy

CDR – Clinical dementia rating

ComBat – Combatting batch effects when combining batches of gene expression microarray data

CSD – Constrained spherical deconvolution

CSF – Cerebrospinal fluid

CVA – Cerebrovascular accident

dMRI – Diffusion magnetic resonance imaging

DTI – Diffusion tensor imaging

DWI – Diffusion-weighted imaging

EDIs – Epidemiology of Dementia in Singapore

FA – Fractional anisotropy

FDR – False discovery rate

FLAIR – Fluid-attenuated inversion recovery

FMRIB – Functional Magnetic Resonance Imaging of the Brain

FN – False negative

FOD – Fibre orientation distribution

FP – False positive

FSL – Functional Magnetic Resonance Imaging of the Brain (FMRIB) software library

GE – Global efficiency

GLM – General linear model

GM – Gray matter

GRAPPA - Generalized autocalibrating partial parallel acquisition

HCP – Human connectome project

ICC – Intraclass correlation coefficient

MD – Mean diffusivity

MMSE – Minimental state examination

MoCA - Montreal Cognitive Assessment

MRI – Magnetic resonance imaging

MST – Minimum spanning tree

NOS – Number of streamlines

PSMD – Peak width of skeletonized mean diffusivity

RISH – Rotation invariant spherical harmonic

ROI – Region of interest

SENSE – Sensitivity encoding

SH – Spherical harmonic

SL – Streamline length

STRIVE - Standards for Reporting Vascular Changes on Neuroimaging

SVD – Small vessel disease

TBSS – Tract-based spatial statistics

TE – Echo time

TI – Inversion time

TIA - Transient ischemic attack

TR – Repetition time

WM – White matter

WMH – White matter hyperintensity

PUBLICATIONS

In this thesis

De Brito Robalo, B. M., Biessels, G.J., Chen, C., Dewenter, A., Duering, M., Hilal, S., Koek, H. L., Kopczak, A., Yin Ka Lam, B., Leemans, A., Mok, V., Onkenhout, L. P., van den Brink, H., & de Luca, A. (2021). *Diffusion MRI harmonization enables joint-analysis of multicentre data of patients with cerebral small vessel disease*.
NeuroImage: Clinical, 32(October); DOI: 10.1016/j.nicl.2021.102886

De Brito Robalo, B. M., Vlegels, N., Leemans, A., Reijmer, Y. D., & Biessels, G. J. (2022). *Impact of thresholding on the consistency and sensitivity of diffusion MRI-based brain networks in patients with cerebral small vessel disease*.
Brain and Behavior, e2523. https://doi.org/10.1002/brb3.2523

De Brito Robalo, B. M., Vlegels, N., Meier, J., Leemans, A., & Biessels, G. J. (2020). *Effect of Fixed-Density Thresholding on Structural Brain Networks : A Demonstration in Cerebral Small Vessel Disease*.
Brain Connectivity.Apr 2020.121-133; DOI: 10.1089/brain.2019.0686

De Brito Robalo, B. M., Biessels, G.J., Chen, C., Dewenter, A., Duering, M., Hilal, S., Koek, H. L., Kopczak, A., Yin Ka Lam, B., Leemans, A., Mok, V., Onkenhout, L. P., van den Brink, H., & de Luca, A. (2022). *Improved Sensitivity and Precision in Multicentre Diffusion MRI Network Analysis Using Thresholding and Harmonization*.
NeuroImage: Clinical, *Accepted*

Not in this thesis

Ferro D., Heinen R, **De Brito Robalo B. M.**, Kuijf H., Biessels G.J., Reijmer Y. (2019). *Cortical microinfarcts and white matter connectivity in memory clinic patients*. Front Neurol 10:571.

Conference abstracts

De Brito Robalo B.M., Vlegels N., Heinen R., Leemans A., Biessels G. J, Reijmer Y.D. C*haracterizing the core network: an unbiased method for structural brain network analysis in small vessel disease*. (VasCog 2018) The 9th International Conference of The International Society of Vascular Behavioural and Cognitive Disorders. Hong Kong.

De Brito Robalo B.M., Vlegels N., Leemans A., Biessels G. J, Reijmer Y.D. F*ixed density thresholding in structural brain networks: a requirement for clinical studies?*. (OHBM 2019)

25th Annual Meeting of the Organization for Human Brain Mapping. Rome

De Brito Robalo, B. M., Jan Biessels, G., Chen, C., Dewenter, A., Duering, M., Hilal, S., Koek, H. L., Kopczak, A., Yin Ka Lam, B., Leemans, A., Mok, V., Onkenhout, L. P., van den Brink, H., & de Luca, A. *Diffusion MRI harmonization enables joint-analysis of multicentre data of patients with cerebral small vessel disease*. (VasCog 2021 Virtual): 11th International Conference of the VasCog Society. September 2021. Newcastle University, UK.

De Brito Robalo, B. M., Jan Biessels, G., Chen, C., Dewenter, A., Duering, M., Hilal, S., Koek, H. L., Kopczak, A., Yin Ka Lam, B., Leemans, A., Mok, V., Onkenhout, L. P., van den Brink, H., & de Luca, A. *Diffusion MRI harmonization and thresholding improve multicentre network analysis: a demonstration in cerebral small vessel disease*. Joint Annual Meeting ISMRM-ESMRMB & ISMRT 31st Annual Meeting May 2022.

Awards

Young investigator award, The 9th International Conference of The International Society of Vascular Behavioural and Cognitive Disorders. Hong Kong.

ACKNOWLEDGEMENTS (DANKWOORD)

I would like to thank everyone who has contributed in any way to this thesis. Without your support and contribution, this thesis would not have been possible. In particular, I would like to thank the participants and patients who contribute with their data for research. Without you, scientific progress is not possible.

Prof. dr. Biessels, dear Geert Jan, it was a real pleasure to complete this PhD trajectory under your supervision. I have always admired your involvement and that you truly care that your students not only understand what they are doing but why they are doing it. I am grateful for the space you have given me to use my creativity and walk my own path, but also for keeping me on track when you saw me "wandering" away from the goal post. I will cherish our conversations and brainstorming sessions, which always left "my brain cooked" as I used to comment with Naomi and Hilde. This is probably because we spoke two different languages, literally (Dutch vs. Portuguese) and figuratively (Doctor vs. Engineer).

Dr. Leemans, dear Alexander, I never mentioned this to you, but before I entered your office for the first time, I had read many of your papers, and I was very impressed (and a bit intimidated) by your knowledge. After the first minutes I realized you were very friendly, and you made me feel welcome in the PROVIDI lab. You were always available for a chat and sometimes even treated us with amazing pooled pork for lunch. I am glad that you guided me during this PhD trajectory and I have learned a lot from you.

Dr. Reijmer, dear Yael, I remember from the moment I had my first interview with you back in May 2017 I knew I wanted to work with you. I was very lucky to have you has a co-promotor for the first half of my PhD. You were always very kind, patient and I look up to you as a researcher at the bridge between clinic and development of new technologies. I have learned a lot from you.

Dr. de Luca, Alberto, my friend. I am extremely lucky to have had you as daily supervisor for the second half of my PhD. We were already friends before that and when I knew you were joining the team, I wondered whether that could affect our work. I couldn't have been more wrong. I would like to thank you for your guidance, support and patience for the many headaches that I probably gave through these past years. I admire your work and learned so much during our meetings, calls and discussions over espressos and cappuccinos. Thank you for everything.

I would also like to thank all co-authors for the great collaboration in all research projects. I was lucky enough to have met and collaborated with researchers from Germany, Singapore and Hong Kong. Special thanks to Prof. Marco Düring, Prof. Christopher Chen, Prof. Christopher Chen and Prof. Vincent Mok.

To my dear colleagues, Hilde, Naomi, Nick and Chloe, thank you for making the work environment so enjoyable. Thanks to you I learned many dutch words and sayings that I will probably

not be able to use in any other context. We all went through the same journey and your support has been crucial. I would also like to thank everyone from the VCI group who made me feel welcome, taught me the "broodje culture" for lunch and made conferences feel like a school trip. Angela, I will do this in dutch since you always insisted that I learn: Bedankt voor je hulp bij alles, voor het feit dat ik me welkom voelde en voor het onmiddellijk oplossen van elk probleem.

Dear paranymphs, Filipe, Naomi and Francisco, what would I have done without you. I thank you for all the help organizing this special day.

Dear Chico, "dude" I know you already for 11 years and it still seems like yesterday that we first met at FCUL and went through Praxe and the crazy years of university. Thank you for your friendship, support, kindness, "memes", movies, games, anime recommendations and NSFW links you send daily to keep the laughter going on. Lieke, thank you for being who you are. I always find it funny that even though you are younger than me it always seems like I come to you for advice, help and support. Do you remember that night when I lost my keys? We will not speak more about it. Thank you for being so kind and carrying, for the good dinners, board games and Mario kart, but also for being so patient with all the jokes me and Chico make constantly. Filipe, I have to thank you again. Nobody else could have the best (?) music taste as you do, knowledge of bars and be so cheerful as you are.

"Namaste boys and girls": Petra, Carlo and Gio thank you for the best dinners ever (my mom has a run for her money), for the trips, Christmas markets and surprise birthday gifts. But most importantly thank you for being there. Dear Biltstraat friends, those were some amazing days: Cinta, Alan, Bea, Vania, Martine, Daniel, Lef, Peggy you really made my stay in Utrecht feel like a second home. Nils, Claudia and Andrea, what would our summers be without the surfing, football and volleyball sessions.

To my in laws, Alie, Gert, Jarmo, Tika and Milan, thank you for your support, kindness and sailing days.

To my family: dear brother, Samuel. You are an example to me, look how far we have come. I am so proud of the person you are, the love you have for your family and how much you care about everyone before yourself. From kids playing street football in our neighborhood to where are now, I'm always looking up to you. To my sisters, Vania and Eveliny, thank your for your love, kindness, funny moments, conversations, even for stealing my toys and desserts when we were kids. To my sister-in-law, Samara and to my nephew Luis, thank you for keeping me grounded. Thank you for being there through everything. Dear cousin Keven, even though we have the whole Atlantic ocean between us, distance means nothing. You have always been close though this all journey and your support means a lot. I'm proud of what you have achieved, and how much you dream for yourself and those around you.

Aos meus pais: palavras não serão capazes de descrever o quanto vocês significam para mim.

Tudo o que conquiestei até agora, a pessoa que sou e onde estou graças a você. Graças ao vosso sacrifício, apoio e amor incondicional. Mãe tu és a pessoa mais forte que eu conheço. Eu sei que te dou muitas dores de cabeça e preocupações por estar tão longe de casa. Preocupas-te com coisas que eu acho tão importantes: se me alimento bem, se me visto bem etc. E tu sempre respondes que um dia eu vou entender quando for a minha vez de cuidar de alguém. Sou teimoso, mas admito que provavelmente tens razão.

Dear Mayke, what an adventure the past year has been. You bring out the best of me. I'm happy I can have you by my side, and I am grateful for your love, kindness, warm heart and for showing me a different perspective of life. You are there through the good and bad moments. I'm looking forward for what the future has in store for us. Life is more beautiful with you!

ABOUT THE AUTHOR

Bruno Robalo was born on May 15th, 1992 in Cape Verde and moved to Portugal later that year. He is the second child of a family of four children, one older brother and two younger sisters. Since a young age, he was always very curious and like to explore the world outside. Eager to learn about science, he followed the Math, Physics and Biology track during his high school years. In 2011 he was accepted for a Bachelor and master's in biomedical engineering and Biophysics at the Faculty of Sciences of the University of Lisbon. His interest in medical technology grew more and more during his university years and he pursued a specialization in radiotherapy and medical imaging. During his Master he did two internships in France (University of Marseille in 2014 and University of Lorraine in 2015-2016). Both internships focused on diffusion MRI, which further in diffusion MRI which sparked is interest in the topic. He graduated with honors in October 2016 and was eager for new adventures.

In 2017 he made the big journey to the rainy Netherlands and was accepted for a PhD in diffusion MRI and brain connectivity in dementia, at the Neurology department of the University Medical Center in Utrecht. This was a project envisioned by Dr. Yael D. Reijmer and Prof. Geert Jan Biessels, who would be his supervisors, together with Dr. A. Leemans from the Radiology department. His PhD journey was full of trials and tribulations, with a change in supervision (with Dr. Alberto de Luca joining the team) and two years of Covid pandemic. Fortunately, the positives far outweighed the negatives, with the support of colleagues, supervisors, family and friends, making the PhD journey a joyful one.

He currently lives in Utrecht, with high hopes for the future.