NeuroImage 195 (2019) 285–299

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms

Chantal MW. Tax [a,*], Francesco Grussu [b,c], Enrico Kaden [c], Lipeng Ning [d], Umesh Rudrapatna [a], C. John Evans [a], Samuel St-Jean [e], Alexander Leemans [e], Simon Koppers [f,g], Dorit Merhof [g], Aurobrata Ghosh [c], Ryutaro Tanno [c,h], Daniel C. Alexander [c], Stefano Zappalà [a], Cyril Charron [a], Slawomir Kusmia [a], David EJ. Linden [a], Derek K. Jones [a,i], Jelle Veraart [j,k]

[a] Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, United Kingdom
[b] Queen Square MS Centre, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, London, United Kingdom
[c] Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom
[d] Harvard Medical School, Boston, MA, United States
[e] Image Sciences Institute, Department of Radiology, University Medical Center Utrecht, Utrecht, the Netherlands
[f] Department of Radiology, University of Pennsylvania and the Children's Hospital of Philadelphia, Philadelphia, PA, United States
[g] Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany
[h] Machine Intelligence and Perception Group, Microsoft Research Cambridge, Cambridge, United Kingdom
[i] Mary McKillop Institute for Health Research, Australian Catholic University, Melbourne, Australia
[j] New York University, New York, NY, United States
[k] imec-Vision Lab, Department of Physics, University of Antwerp, Antwerp, Belgium

A B S T R A C T

Diffusion MRI is being used increasingly in studies of the brain and other parts of the body for its ability to provide quantitative measures that are sensitive to changes in tissue microstructure. However, inter-scanner and inter-protocol differences are known to induce significant measurement variability, which in turn jeopardises the ability to obtain 'truly quantitative measures' and challenges the reliable combination of different datasets. Combining datasets from different scanners and/or acquired at different time points could dramatically increase the statistical power of clinical studies, and facilitate multi-centre research. Even though careful harmonisation of acquisition parameters can reduce variability, inter-protocol differences become almost inevitable with improvements in hardware and sequence design over time, even within a site. In this work, we present a benchmark diffusion MRI database of the same subjects acquired on three distinct scanners with different maximum gradient strength (40, 80, and 300 mT/m), and with 'standard' and 'state-of-the-art' protocols, where the latter have higher spatial and angular resolution. The dataset serves as a useful testbed for method development in cross-scanner/cross-protocol diffusion MRI harmonisation and quality enhancement. Using the database, we compare the performance of five different methods for estimating mappings between the scanners and protocols. The results show that cross-scanner harmonisation of single-shell diffusion data sets can reduce the variability between scanners, and highlight the promises and shortcomings of today's data harmonisation techniques.

1. Introduction

Diffusion-weighted magnetic resonance imaging (dMRI) is being used increasingly to characterise tissue microstructure in health and disease (Johansen-Berg and Behrens, 2009; Jones, 2010a). Despite the promise of dMRI providing quantitative measures related to tissue microstructure, an inherent variability exists in the measurements when the same experiment is repeated on different scanners or at different time points. This inter- and intra-scanner variability can be caused by various factors including, but not limited to, differences in field strength, maximum available gradient strength, reconstruction technique from k-space data, positioning of the participant, imaging gradient non-linearities, number and sensitivity of the receiver coils, software versions used, and changes in the system calibration (Mirzaalian et al., 2016).

In addition to scanner-related variations, differences in acquisition protocol parameters introduce an extra source of variability in the measurements (Jones, 2010b). For example, even though guidelines for diffusion tensor MR imaging (DT-MRI) acquisitions have been proposed (e.g., Jones and Leemans, 2011), a standardised protocol is currently missing, i.e. the number and distribution of diffusion gradient directions

* Corresponding author.
E-mail address: taxc@cardiff.ac.uk (C.MW. Tax).

https://doi.org/10.1016/j.neuroimage.2019.01.077
Received 15 August 2018; Received in revised form 16 January 2019; Accepted 30 January 2019
Available online 1 February 2019
1053-8119/

in clinical research protocols tend to vary across sites. Nevertheless, protocol differences in studies that involve multiple scanners and/or are done over long periods of time are sometimes inevitable and not necessarily a sign of suboptimal experiment design. For example, developments of biophysical models increased the adoption of multiple b-value diffusion acquisition protocols over time, which concomitantly further increases the degrees of freedom in protocol design. Moreover, with technical advances in scanner hardware and software, protocol-updates during a study become desirable because they allow investigators to exploit such technical improvements in data acquisition. Notably, simultaneous multislice imaging (Feinberg et al., 2010; Larkman et al., 2001; Nunes et al., 2006) allows for the acquisition of more image volumes per unit time, and stronger-gradient systems (Jones et al., 2018; Setsompop et al., 2013) facilitate the acquisition of higher SNR data per unit time because of the reduced echo time (TE), where a trade-off can be made with smaller voxel-sizes and/or higher b-values.

Notwithstanding the challenges introduced by cross-scanner and cross-protocol differences, in the current era of "big data" there is strong interest in reliable combination of data acquired on different MRI scanners and/or with different protocols. Combining data from different scanners could increase the statistical power and sensitivity of studies, with obvious benefits in trials and multi-centre research, particularly in rare diseases or with difficult-to-recruit participants. Combining data from different protocols and quality could enable the transfer of rich information content from state-of-the-art acquisitions (e.g. from specialized systems as the 300 mT/m gradient Connectom system (Setsompop et al., 2013)), to lower quality data, e.g. to enhance spatial or angular resolution, or to enhance features that are less pronounced in low b-value measurements (Alexander et al., 2017; Jones et al., 2018).

The process of finding a mapping between diffusion data sets acquired with different scanners or protocols and making them as comparable as possible has gained increased attention recently (Fortin et al., 2017; Karayumak et al., 2018; Mirzaalian et al., 2017, 2016; Pohl et al., 2016). Often, these approaches are evaluated on databases from different scanners where the subjects are matched for age, gender, handedness, and socio-economic status, such that no statistical differences are expected at the group level. Ideally, however, individual subjects would be rescanned on different systems in relatively quick succession, such that measurement differences can be clearly attributed to inter-scanner and/or inter-protocol differences. Such databases have been acquired in the context of testing the reproducibility of DT-MRI metrics across scanners with different field strengths (1.5T vs 3T), sites, software versions, and vendors (Grech-Sollars et al., 2015; Vollmar et al., 2010; Zhu et al., 2011), but are (to the best of our knowledge) not publicly available.

Here, we present a benchmarking database of human brains that provides a testbed for data harmonisation[1] across 3 different scanners and 5 different acquisition protocols, along with a comparison of 5 dMRI harmonisation algorithms. The database consists of acquisitions of the same 14 healthy participants scanned on MR systems with different maximum gradient strength (40, 80, and 300 mT/m). On the 80 mT/m and 300 mT/m systems, two types of protocols were acquired: 1) a 'standard' protocol with acquisition parameters matched as closely as possible to those on the 40 mT/m system (i.e. a typical clinical protocol); and 2) a 'state-of-the-art' protocol where the superior hardware and software specifications were utilised to increase the number of acquisitions and spatial resolution per unit time. In a recent open competition[2], entrants were invited to implement and optimise algorithms that would harmonise data collected under these different scenarios. The algorithms are evaluated based on their performance in two tasks: 1) matched resolution scanner-to-scanner mapping between the standard acquisitions; and 2) spatial and angular resolution enhancement, finding a mapping between the standard acquisition of the 40 mT/m system to the state-of-the-art acquisition of the other systems.

2. Methods

The database and acquisition parameters for the three different scanners are described in detail in Section 2.1. Section 2.2 describes the harmonisation tasks that were formulated for the open competition where entrants were invited to evaluate their algorithms on the database. To minimise confounding effects of differences in preprocessing between different harmonisation algorithms during evaluation, we provided a minimally processed version of the database (described in Section 2.3). Section 2.4 describes the strategy used to evaluate algorithm outputs, and Section 2.5 summarises the algorithms proposed for scanner-to-scanner mapping and image quality enhancement to solve the proposed inter-scanner mapping tasks.

2.1. Data

14 healthy volunteers were included in the study (10 females, average age 25.7 years with range 21–41 years, Table 1), which was approved by Cardiff University School of Psychology ethics committee. Written informed consent was obtained from all subjects. The same 14 subjects were scanned on three different 3T scanners with different maximum gradient strengths: a) 3T GE Signa Excite HDx (40 mT/m), b) 3T Siemens Prisma (80 mT/m), and c) 3T Siemens Connectom (300 mT/m). The average time between acquisitions on scanners a) and b), and a) and c) was 21 and 22 months, respectively. The scanners had no software upgrades during the course of the study.

Spin-echo echo-planar dMRI images (SE-EPI) were acquired with a 'standard' (ST) protocol on all three scanners, and a 'state-of-the-art' (SA) protocol on the 80 mT/m and 300 mT/m systems (Table 2, Fig. 1). For the SA protocol, we exploited multiband-acquisition and the stronger gradients to shorten TE and improve the spatial- and angular resolution per unit time. Additional $b = 0\,s/mm^2$ images were acquired with TE and/or TR matching between protocols. Magnitude data was obtained for all scanners and protocols, and phase data was additionally saved for the 80 mT/m and 300 mT/m systems. Further information on the estimated SNR (Veraart et al., 2016b) is reported in Supplementary Table 1. Structural MPRAGEs (Magnetization Prepared RApid Gradient Echo) (de Lange et al., 1991) were acquired for each scanner and subject. The data is available for researchers upon request (see Section 4.5).

Table 1
Healthy volunteers included in the study.

Subject	Age[a]	Gender	Time gap scans a) and b) [months]	Time gap scans a) and c) [months]
A	26	f	14	16
B	21	m	23	24
C	41	f	14	16
D	25	f	23	23
E	21	f	21	21
F	25	f	28	30
G	25	f	28	28
H	28	f	20	20
I	21	m	21	21
J	22	f	23	23
K	26	m	28	30
L	35	m	16	16
M	23	f	23	23
N	21	f	20	20

[a] Age at the time of the first scan.

[1] Here we use the word 'harmonisation' in the context of finding a mapping between datasets, irrespective of their difference in quality.

[2] These data were collected for the 'Diffusion MRI Harmonisation Challenge', an initiative presented at the 2017 Computational Diffusion MRI Workshop of the MICCAI conference in Quebec City, Canada https://projects.iq.harvard.edu/cdmri2017/home.

Table 2
Acquisition parameters for the different scanners and protocols.

Scanner	GE 40 mT/m	Siemens 80 mT/m		Siemens 300 mT/m	
Protocol	Standard (ST)	Standard (ST)	State-of-the-art (SA)	Standard (ST)	State-of-the-art (SA)
Diffusion weighted images					
Sequence	TRSE	PGSE	PGSE	PGSE	PGSE
b-values [s/mm^2]	1200	1200, 3000	1200, 3000, 5000	1200, 3000	1200, 3000, 5000
# directions per b-value	30	30	60	30	60
TE [ms]	89	89	80	89	68
TR [ms]	Cardiac gated	7200	4500	7200	5400
Δ /δ [ms]		41.4/26.0	38.3/19.5	41.8/28.5	31.1/8.5
$\delta_1 = \delta_4/\delta_2 = \delta_3$ [ms]	11.23/17.84				
Phase encoding direction	AP	AP	AP	AP	AP
Acquired voxel size [mm^3]	2.4 × 2.4 × 2.4	2.4 × 2.4 × 2.4	1.5 × 1.5 × 1.5	2.4 × 2.4 × 2.4	1.2 × 1.2 × 1.2
Reconstructed voxel size	1.8 × 1.8 × 2.4	1.8 × 1.8 × 2.4	1.5 × 1.5 × 1.5	1.8 × 1.8 × 2.4	1.2 × 1.2 × 1.2
Matrix size	96 × 96	96 × 96	154 × 154	96 × 96	180 × 180
# slices	60	60	84	60	90[a]
SMS factor	1	1	3	1	2
Parallel imaging	ASSET 2	GRAPPA 2	GRAPPA 2	GRAPPA 2	GRAPPA 2
Bandwidth [Hz/Px]	3906	2004	1476	2004	1544
Partial Fourier	5/6	–	6/8	6/8	6/8
Coil combine		Adaptive combine	Sum of Squares[b]	Adaptive combine	Adaptive combine
Head coil	8 channel	32 channel	32 channel	32 channel	32 channel
b0 images					
TE [ms]	89	89, 80, 89	80, 80, 89	89, 68, 89	68, 68, 89
TR [ms]	Cardiac gated	7200, 7200, 13000	4500, 7200, 7200	7200, 7200, 13000	5400, 7200, 7200
Phase encoding direction	AP	AP, PA	AP, PA	AP, PA	AP, PA

[a] A trade-off was made between number of slices and timing parameters, as a result the cerebellum was not always covered.
[b] This was the only strategy possible with the settings used. TRSE = twice-refocused spin-echo, PGSE = pulsed-gradient spin-echo.

2.2. Harmonisation tasks

The rich database, which includes images of the same subject acquired on different scanners with different b-values, angular resolutions, spatial resolutions, and timing parameters, allows for a wide variety of aspects to be evaluated. In this work, we focus on evaluating the process of finding a mapping between the lowest b-value shells across scanners and protocols. The b = 0 and 1200 s/mm^2 data of all scanners and protocols from 10 randomly selected subjects (coined A-G, I-K in this study) were used as training dataset. Data from the remaining 4 subjects (H, L-N) were used as a testing set; the 40 mT/m data were distributed to the entrants of the challenge while the data of the other scanners were held back for further evaluation purposes.

Two tasks were evaluated:

1) Scanner-to-scanner mapping at matched resolution acquisition protocol, predicting the 80 mT/m and 300 mT/m ST signals in the cerebrum from the 40 mT/m ST signals; and
2) Spatial- and angular resolution enhancement, predicting the 80 mT/m and 300 mT/m SA signals in the cerebrum from the 40 mT/m ST signals.

2.3. Preprocessing

All datasets were manually checked for artifacts such as slice outliers, vibration artifacts, and interleave motion artifacts (Gallichan et al., 2009; Tax et al., 2016; Tournier et al., 2011). One DWI volume was excluded from one dataset (subject H of the test set). The data were subsequently preprocessed for each subject and scanner as detailed in the next paragraphs, where the steps were homogenised where possible and an additional gradient-nonlinearity distortion correction step was added for the 300 mT/m system. All data was spatially registered for each subject across scanners (with the mean DWI of the 80 mT/m ST acquisition as template), and the result was inspected manually.

The 40 mT/m data were corrected for eddy current distortions and subject motion with FSL EDDY (Andersson and Sotiropoulos, 2016) and corrected for EPI distortions by nonlinear registration of the mean of the DWIs to the 80 mT/m ST mean DWI with Elastix (Irfanoglu et al., 2015; Klein et al., 2010; Leemans et al., 2009) and b-matrix rotation (Leemans and Jones, 2009).

The 80 mT/m data were corrected for eddy current distortions, subject motion, and EPI distortions with FSL TOPUP (Andersson et al., 2003) and EDDY. The corrected 80 mT/m SA mean DWI was affinely registered to the 80 mT/m ST mean DWI with b-matrix rotation.

The 300 mT/m data were corrected for eddy current distortions, subject motion, EPI distortions, and gradient-nonlinearity distortions (Glasser et al., 2013) with FSL TOPUP and EDDY and in-house software kindly provided by Martinos Centre, Massachusetts General Hospital. The corrected 300 mT/m ST and SA mean DWI were affinely registered to the 80 mT/m ST mean DWI with b-matrix rotation.

The MPRAGE of each scanner was affinely registered to the 80 mT/m ST mean DWI. Face removal was subsequently performed (Bischoff--Grethe et al., 2007). Brain masks excluding the cerebellum were obtained from the MNI atlas, by warping the mask in MNI space non-linearly to each subject's DWI space with a repeated call of FSL FNIRT and affine registration from FSL FLIRT used for initialisation (Andersson et al., 2007; Jenkinson et al., 2012).

2.4. Evaluation

Harmonisation algorithms had to predict the image matrix of the preprocessed 80 mT/m and 300 mT/m datasets by using the b-matrix files and the associated 40 mT/m image data of each of the four test subjects. From the *predicted* data of each test subject and each algorithm, the diffusion tensor was estimated using a weighted linear least squares estimator (Veraart et al., 2013) using the MRTrix software package, and fractional anisotropy (FA) and mean diffusivity (MD) were subsequently computed in each voxel. In addition, rotationally invariant spherical harmonic (RISH) features R0 and R2 (Mirzaalian et al., 2015) were computed after normalising the signal per voxel with the mean b = 0 signal to measure the angular frequency of the diffusion signals.

These results were evaluated against the ground truth features derived from the *acquired* data (Fig. 2). Different errors were computed to enable the characterisation of accuracy and precision: mean-error (ME, (predicted - acquired), measuring accuracy), mean-normalised error (MNE, (predicted - acquired)/acquired, measuring relative accuracy),

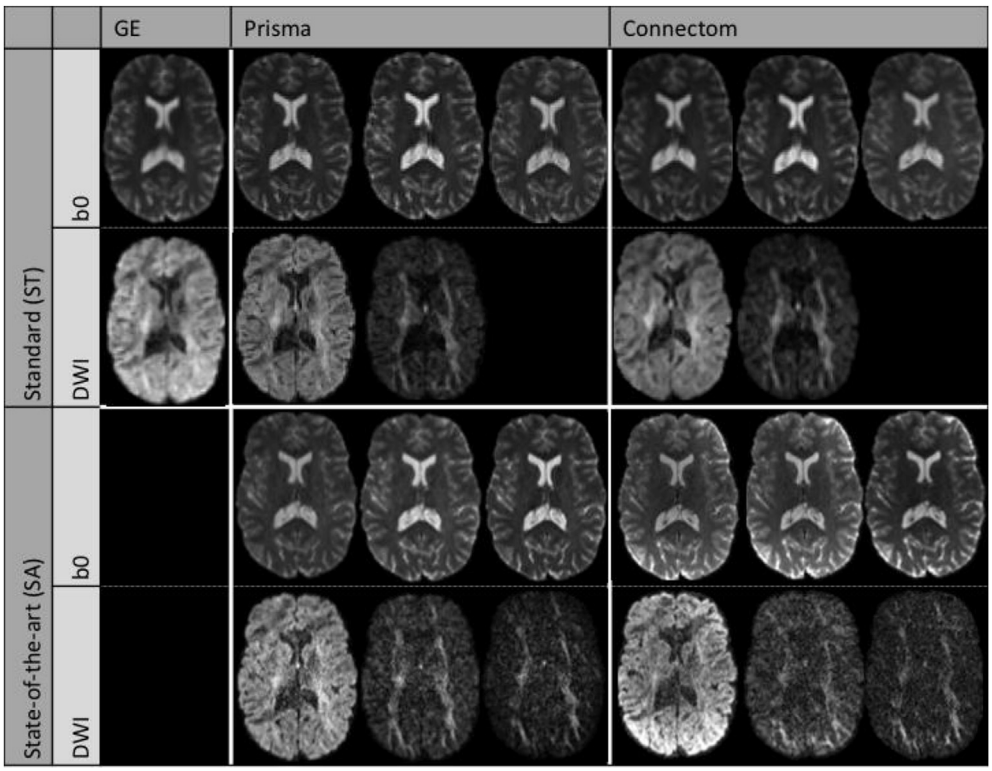

Fig. 1. Example diffusion images of the ST and SA protocols from one subject after preprocessing. b0 images with three different combinations of TE/TR for the 80 mT/m and 300 mT/m scanners. DWIs with b-values (from left to right) 1200, 3000, and 5000 s/mm^2.

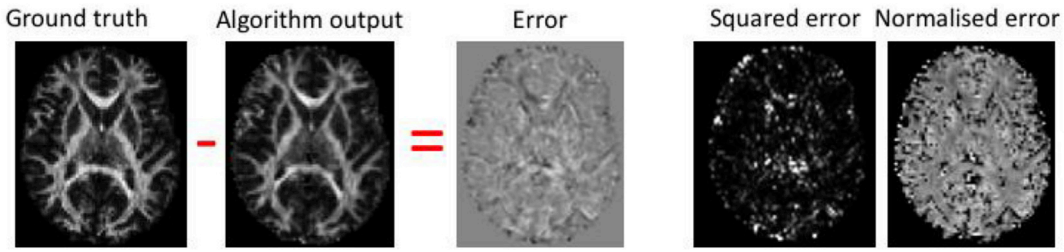

Fig. 2. Evaluation procedure: the error is computed as the difference between the ground truth and predicted image, and from this the squared- and normalised error are computed.

and mean-squared error (MSE, (predicted - acquired)^2, measuring accuracy and precision). The accuracy characterised by ME and MNE can be both positive and negative as this would indicate over/under-estimation of the metric, whereas MSE takes the absolute error into account. These errors were computed globally (in a brain mask), regionally (Freesurfer regions (Fischl et al., 2002)) and locally (sliding $3 \times 3 \times 3$ voxels-neighbourhood). Voxels at the edge of the brain, the cerebellum, and systematically poor performing regions (see Section 3.1.2) were excluded. Further information on the number of brain voxels excluding the cerebellum is reported in Supplementary Table 2.

2.5. Algorithms

Entrants to the open competition developed 5 different algorithms to solve the inter-scanner mapping tasks described in section 2.2 (task 1: scanner-to-scanner mapping of matched acquisition protocols; task 2: spatial/angular resolution enhancement). Additionally, a 'reference' prediction of the 80 mT/m and 300 mT/m ST and SA data was created from the 40 mT/m ST data using simple trilinear interpolation in the spatial domain, and spherical harmonics interpolation (order 6 for ST and 8 for SA) in the angular domain.

The algorithms developed by the entrants explored different deep learning architectures as well as dictionary learning techniques to solve the proposed tasks. They will be referred to as: spherical harmonic network (SHNet); spherical harmonic residual network (SHResNet); spherical network (SphericalNet); sparse dictionary learning (SDL) and fully convolutional shuffling network (FCSNet). A summary of the algorithms is presented in Table 3, a detailed description of each algorithm is provided in the subsections below.

2.5.1. Spherical harmonic network (SHNet)

The SHNet algorithm is a deep learning network inspired by elements of Golkov et al. (2015) and Koppers et al. (2017). In this algorithm, every signal was preprocessed by dividing by its baseline $b = 0$ measurement, followed by a conversion into the SH space (order four and Laplace-Beltrami regularization of $\lambda = 0.006$). The network consisted of three fully connected layers with rectified linear units (ReLU) as activation function, followed by a batch normalization layer to stabilise the training process. An overview is given in Table 4.

For training of the hyperparameters, 9 out of 10 subjects from the training set were used, with the remaining subject used for training validation before deployment on the test set of 4 additional subjects.

Table 3
Summary of harmonisation algorithms evaluated.

Algorithm name	Additional preprocessing	Training domain	Core method	Algorithm details
SHNet	Brain extraction	SH	Deep learning	qDL inspired network, anatomically constrained training
SHResNet	Brain extraction	SH	Deep learning	Residual structure network, anatomically constrained training
SphericalNet	Brain extraction	SH	Deep learning	Local Spherical Convolution Network, anatomically constrained training
FCSNet	Brain extraction	SH	Deep learning	Fully convolutional network with task-dependent regularization (L2 and L1)
SDL		Over-complete data-driven dictionaries	Adaptive dictionary learning	Linear mapping between over-complete dictionaries

Table 4
Topology of SHNet.

#Layer	Type	Parameters	Activation
1	Batch Normalization	–	–
2	Fully-Connected	From: #SH coefficients To: 150 neurons	ReLU
3	Batch Normalization	–	–
4	Fully-Connected	From: 150 neurons To: 150 neurons	ReLU
5	Batch Normalization	–	–
6	Fully-Connected	From: 150 neurons To: 150 neurons	ReLU
7	Batch Normalization	–	–
8	Fully-Connected	From: 150 neurons To: #SH coefficients	–

During training, parameters were initialised using the Adam optimiser (learning rate of 0.001; batch size 128) (Kingma and Ba, 2014), which was replaced by the SGD optimiser (Robbins and Monro, 1951) after the first five epochs. Afterwards, the learning rate was decreased by ten percent if the performance did not improve for more than five epochs within the validation subject. The reduced learning rate leads to a better fine tuning of the network. Furthermore, training was only performed on voxels within a brain mask derived from FSL BET (Smith, 2002).

The SHNet described above was employed for task 1 (matched resolution scanner-to-scanner mapping). For task 2 (spatial/angular resolution enhancement), standard cubic interpolation is also utilised to increase the spatial resolution, while gradients are resampled utilising the predicted SH coefficients to deal with the increased higher angular resolution. The deep learning framework is based on PyTorch. The runtime per voxel on a Nvidia Geforce 1080Ti with 11 GB RAM was 6.4e-05 s.

2.5.2. Spherical harmonic residual network (SHResNet)

The SHResNet algorithm (Koppers et al., 2018) is a deep learning network structure based on the novel concept of residual structure (He et al., 2016), which introduces a subtraction path from the input to the network output, resulting in very robust performance. Furthermore, the network focuses only on the difference between the input and its corresponding target signal. In addition, residual structures are efficiently trainable, even for very deep networks.

Data preprocessing utilised SH (order four and Laplace-Beltrami regularization of $\lambda = 0.006$), while each signal was divided by its $b = 0$ measurement. In this network, three main 3D-convolutional layers (kernel size $3 \times 3 \times 3$) processed the signal and predicted the difference for a specific harmonic order. The first two convolutional layers padded the signal to keep the spatial dimensions, while the last convolutional layer reduced the signal from a $3 \times 3 \times 3$ voxel neighbourhood to a single voxel. Since each SH order was predicted separately, three individual networks were required for an SH order of four, which were combined with a fully connected layer. Afterwards, the resulting signal was subtracted from the corresponding input signal. A final fully connected layer was utilised to smooth the signal and to generate the predicted SH coefficients. An overview is given in Fig. 3.

Similarly to SHNet, SHResNet relied on 9 out of 10 subjects from the training set for the actual training, with the remaining subject used for training validation before deployment on the test set, while only voxel within a brain mask based on FSL BET (Smith, 2002) are considered for training. The network was initialised utilising the Adam optimiser (learning rate of 0.001; batch size 128) (Kingma and Ba, 2014), which is replaced by an SGD optimiser after the first five epochs. After this change, the learning rate was decreased by ten percent if the performance did not improve for more than five epochs.

The SHResNet described above was employed for task 1 (matched resolution scanner-to-scanner mapping). For task 2 (spatial/angular resolution enhancement), standard cubic interpolation is also utilised, as described for SHNet. The deep learning framework is based on PyTorch. The runtime per voxel on a Nvidia GeForce 1080Ti with 11 GB RAM was 0.0014 s.

2.5.3. Spherical network (SphericalNet)

The SphericalNet algorithm utilises a novel deep learning structure based on spherical surface convolutions (Koppers and Merhof, 2018), which were designed especially for spherical signals. These convolutions utilise local gradient neighbourhood information to increase the accuracy of reconstruction, while spatial neighbourhood information is passed layer by layer. In the end, spatial information is combined within the last convolutional layer to project from a $3 \times 3 \times 3$ voxel neighbourhood onto a $1 \times 1 \times 1$ target voxel.

As preprocessing, every signal was transformed into the SH space to avoid a gradient-based mismatching, while the SphericalNet transformed every signal back into a predefined signal space, consisting of 30 equidistantly sampled gradient directions over a hemisphere. Afterwards, every voxel was processed by three spherical convolutions with a kernel size of one plus five and an angular distance of $\Theta = \frac{\pi}{10}$. The angular distance defines the angle between the resulting (in this case five) sampled gradient directions and their corresponding main gradient direction. After each spherical convolution, sigmoid functions were used as activation functions to limit every signal's range between 0 and 1. Subsequently, every signal was converted back into SH space. A batch normalization layer normalised the resulting coefficients. In the end, three 3-D convolutional layers exploited additional spatial neighbourhood information with parametric rectified linear units ($f(x) = k \max(0, x)$, with k being a learnable parameter) (PReLU) as activation functions. A complete overview of the network's architecture is given in Table 5.

The SphericalNet is trained on brain voxels (derived with FSL BET (Smith, 2002)) from 9 out of 10 training set subjects, with the remaining subject used for training validation before deployment on the test set. The network was initialised utilising the Adam optimiser (learning rate of 0.001; batch size 128) (Kingma and Ba, 2014), which is replaced by an SGD optimiser after the first five epochs. After this change, the learning

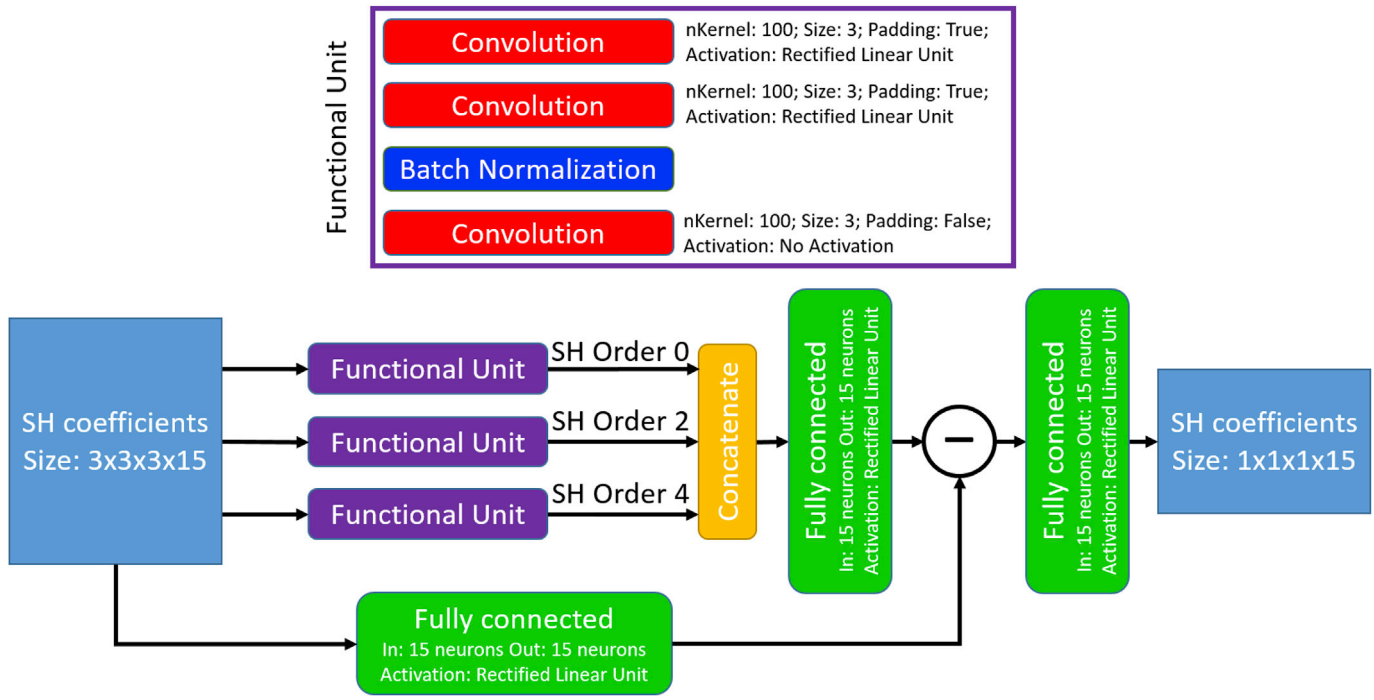

Fig. 3. Structure of the SHRestNet.

Table 5
Architecture of the SphericalNet.

#Layer	Type	Parameters	Activation
1	Conversion	From SH to signal space (SH order 4; Laplace-Beltrami Regularization 0.006)	–
2	Spherical Surface Convolution (applied on gradient signals)	Input: 1 Shell; Output: 16 Shells; kernel size: 5; $\Theta = \frac{\pi}{10}$	Sigmoid
3	Spherical Surface Convolution (applied on gradient signals)	Input: 16 Shell; Output: 16 Shells; kernel size: 5; $\Theta = \frac{\pi}{10}$	Sigmoid
4	Spherical Surface Convolution applied on gradient signals)	Input: 16 Shell; Output: 16 Shells; kernel size: 5; $\Theta = \frac{\pi}{10}$	Sigmoid
5	Conversion	From signal to SH space (SH order 4; Laplace-Beltrami Regularization 0.006)	
6	Batch Normalization	–	–
7	3D Spatial Convolution	Kernel size: $3 \times 3 \times 3$, padding: 1	PReLU
8	3D Spatial Convolution	Kernel size: $3 \times 3 \times 3$, padding: 1	PReLU
9	3D Spatial Convolution	Kernel size: $3 \times 3 \times 3$, padding: 0	

rate was decreased by ten percent if the performance did not improve over five epochs to ensure a good fine tuning of the network.

The SphericalNet described above was employed for task 1 (scanner-to-scanner mapping). For task 2 (spatial/angular resolution enhancement), standard cubic interpolation is also utilised, as described for SHNet. The deep learning framework is based on PyTorch. The runtime per voxel on a Nvidia GeForce 1080Ti with 11 GB RAM was 0.0067 s.

2.5.4. *Fully-convolutional shuffling network (FCSNet)*

The FCSNet algorithm relies on a patch-based fully-convolutional network (FCN) to solve matched resolution scanner-to-scanner harmonisation and resolution enhancement tasks as presented in this paper.

The FCSNet architecture was inspired by Tanno et al. (2017), and contained a "shuffle" operation in the last layer of a super-resolution network to efficiently compute a transpose-convolution as a final step (Shi et al., 2016). The FCSNet structure used here differed from the previous implementation in Tanno et al. (2017) as it contained four hidden layers and a skip connection (Fig. 4). Also, it relied on a different loss function (see below).

FCSNet processed SH coefficients obtained from signals within a brain mask, which was derived from FSL BET (Smith, 2002) and eroded to exclude boundary voxels with noisy signal. SH coefficients were estimated using Dipy (Garyfallidis et al., 2014), up to order 6 for ST protocols and up to order 8 for the SA protocols. For the actual training, SH coefficients were clipped to the 98th percentile of function values on the sphere over the masked image, as this further reduced noise.

For the training processes, sizes of $11 \times 11 \times 11$ and of $3 \times 3 \times 3$ were used for input and output patches respectively, with the number of input and output channels being 29 (input; SH order-6 plus b = 0) and 29 or 46 (output; SH order=6 or 8 plus b = 0). Each hidden layer consisted of a $3 \times 3 \times 3$ convolution layer, a ReLu activation with varying filter lengths and a dropout layer with 0.5 keep probability. The last hidden layer had a skip-connection to the input layer to "sharpen" the prediction. The output layer was computed after a bottleneck convolution (see Fig. 4).

The loss function was constituted of two parts: a channel-wise loss, which gives equal weight to all channels and a loss on the function-value on the sphere, which enforces signal fidelity along hundred uniformly chosen directions on a hemisphere, while also considering the RISH constraints of order 0, 2, 4, 6 for the SA protocol (300 mT/m scanner).

At prediction time, the FCN was applied to juxtaposed input patches to fully cover the subject brains.

Hyperparameters (learning rate, number of layers and batch size) were selected based on the MSE on the validation set. The deep learning framework is based on TensorFlow. Training comprised of 200 epochs consisting of 50000 3D patch-pairs (input size: $11 \times 11 \times 11 \times 29$) selected randomly from the training images and evaluated using mini-batches of size 20 and a learning rate of 1e-4. Training time was in the order of ten hours, while inference time for a masked brain (HCP resolution) is in the order of a minute. All training and testing were done on a server with 112 GB RAM with an NVIDIA GTX Titan X GPU.

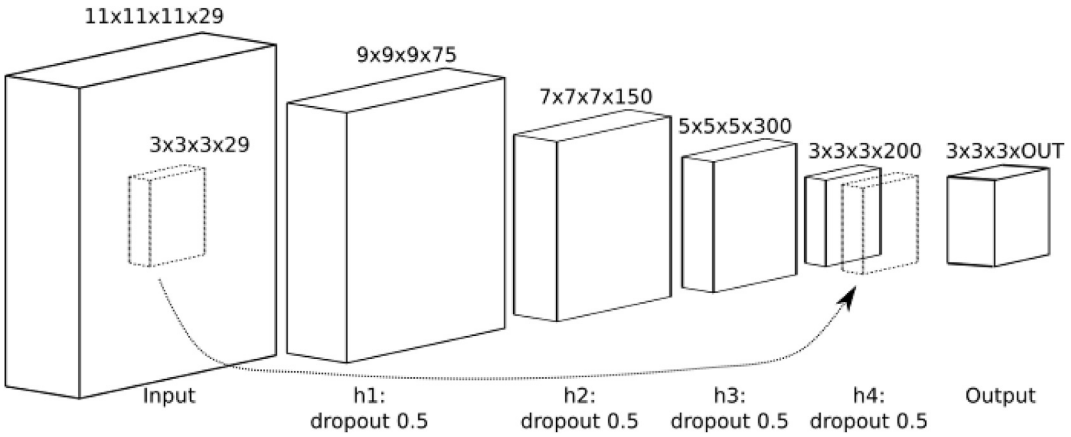

Fig. 4. Architecture of the FCSNet with four hidden layers and a skip connection.

2.5.5. Sparse dictionary learning (SDL)

The SDL algorithm relies on the methodology recently developed in St-Jean et al. (2016, 2017), based on over-complete sparse dictionaries that are learned automatically from the data.

For harmonisation with SDL, N patches of small spatial and angular local neighbourhoods were extracted from all datasets and organised to create a set of column arrays $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, with each $\mathbf{X}_n \in \mathrm{R}^{m\times 1}$. Subsequently, sparse features were automatically created from the target scanner datasets using dictionary learning (Mairal et al., 2010). A sparse dictionary $\mathbf{D} \in \mathrm{R}^{m\times p}$ was found such that

$$\mathbf{D}^* = \mathrm{argmin}_{\mathbf{D}} \sum_{n=1}^{N} ||\mathbf{X}_n - \mathbf{D}\boldsymbol{\alpha}_\mathbf{n}||_2^2 + \lambda||\boldsymbol{\alpha}_n||_1, s.t. ||\mathbf{D}||_2^2 = 1, \alpha \geq 0, \quad (1)$$

with $\boldsymbol{\alpha}_n \in \mathrm{R}^{p\times 1}$ being an array of non-negative coefficients and $\mathbf{D}$ the dictionary initialised from patches randomly extracted from the datasets, set to have twice as many columns as rows (i.e. $p = 2m$). Iterative updates alternating between refining $\mathbf{D}$ using eq. (1) (and holding $\boldsymbol{\alpha}$ fixed) and updating $\boldsymbol{\alpha}$ (with $\mathbf{D}$ held fixed) with a coordinate descent scheme (Friedman et al., 2010) were carried for 1000 iterations using a batchsize of 128 patches randomly sampled for each iteration. An automatic search for the regularization parameter λ was employed (Friedman et al., 2010). The search selected the value of λ according to the Akaike Information Criterion (AIC), where the number of non-zero elements in the dictionary was used as the number of degrees of freedom for the model (Tibshirani and Taylor, 2012).

For the reconstruction in task 1 (matched resolution scanner-to-scanner mapping), the dictionary was created using patches of size $3 \times 3 \times 3 \times 5$. Images were mean-subtracted and standardised to account for scaling, with the coefficients $\boldsymbol{\alpha}_n$ subsequently unscaled afterwards. This idea assumes that there is a set of common features which can be mapped between acquisitions made on different scanners. The general reconstruction process is shown in Fig. 5.

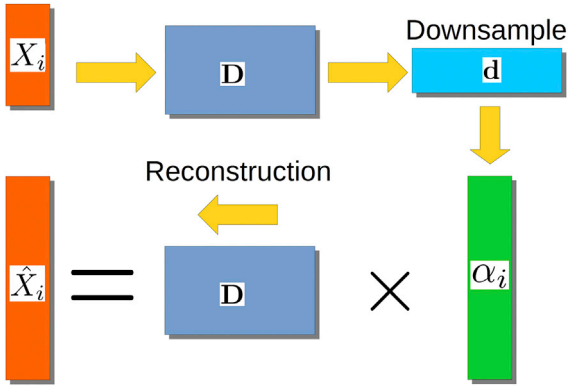

Fig. 5. Reconstruction process for SDL. Local patches are decomposed into vectors $\mathbf{X}_i$ to build the dictionary $\mathbf{D}$. From there, the coefficients α_i are computed from $\mathbf{D}$ for task no. 1 or using a downsampled version of $\mathbf{D}$ for task no. 2. The final reconstruction for each patch $\mathbf{X}_i$ is obtained by multiplying $\mathbf{D}$ and α_i.

For the reconstruction in task 2 (spatial and angular resolution enhancement), patches of different spatial sizes were extracted from the images at lower resolution (ST protocol; patches of sizes $3 \times 3 \times 3$) and from the images at higher resolution (SA protocol; patches of size $5 \times 5 \times 5$ and $6 \times 6 \times 6$), under the hypothesis that such sizes would yield a plausible representation between the lower resolution and higher resolution scans. Reconstruction coefficients $\boldsymbol{\alpha}_n$ were computed on the downsampled dictionary and the final reconstruction used the original size dictionary (St-Jean et al., 2017). Finally, to match the gradient directions, the truncated SH basis of order 6 (Descoteaux et al., 2007) was used on each final dataset to predict the target images at the required gradient directions.

The training time for the matched resolution scanner-to-scanner mapping within a brain mask, 1000 epochs, was approximately 72 min on a quad cores Intel Xeon processor at 3.5 GHz with an average ram usage of around 600 MB. For predicting each dataset from the 40 mT/m scanner to the target scanner, it took approximately 4h30 min per dataset with an approximate ram usage of 315 MB.

3. Results

3.1. Matched resolution scanner-to-scanner mapping

3.1.1. Global evaluation

Fig. 6 shows the global MSE of FA, MD, R0, and R2, respectively, i.e. the mean taken over all voxels within the mask. Results are shown for the 4 different test subjects. There does not seem to be a systematic deviation for one of the subjects, so all subjects were included for evaluation. Global MSE are shown for the different harmonisation approaches described in Section 2.5 (SHNet, SHResNet, SphericalNet, FCSNet, and SDL) along with the 'reference' prediction obtained from trilinear interpolation in the spatial domain and spherical harmonics interpolation in the angular domain.

In most cases, the MSE for the harmonisation methods were lower than the reference. For FA, MD, and R2, this was true both for the 80 mT/m and 300 mT/m predictions. For R0, the MSE of the harmonisation methods was higher than that of the reference for the 80 mT/m prediction, except for subject H. For this subject, the SHNet 300 mT/m prediction has a consistently higher MSE than the reference for FA, MD, and R2. Fig. 6, bottom also shows the MSE of the submissions compared to scan-rescan differences, which provide the ultimate goal for data harmonisation. Scan-rescan data was only available for the 300 mT/m

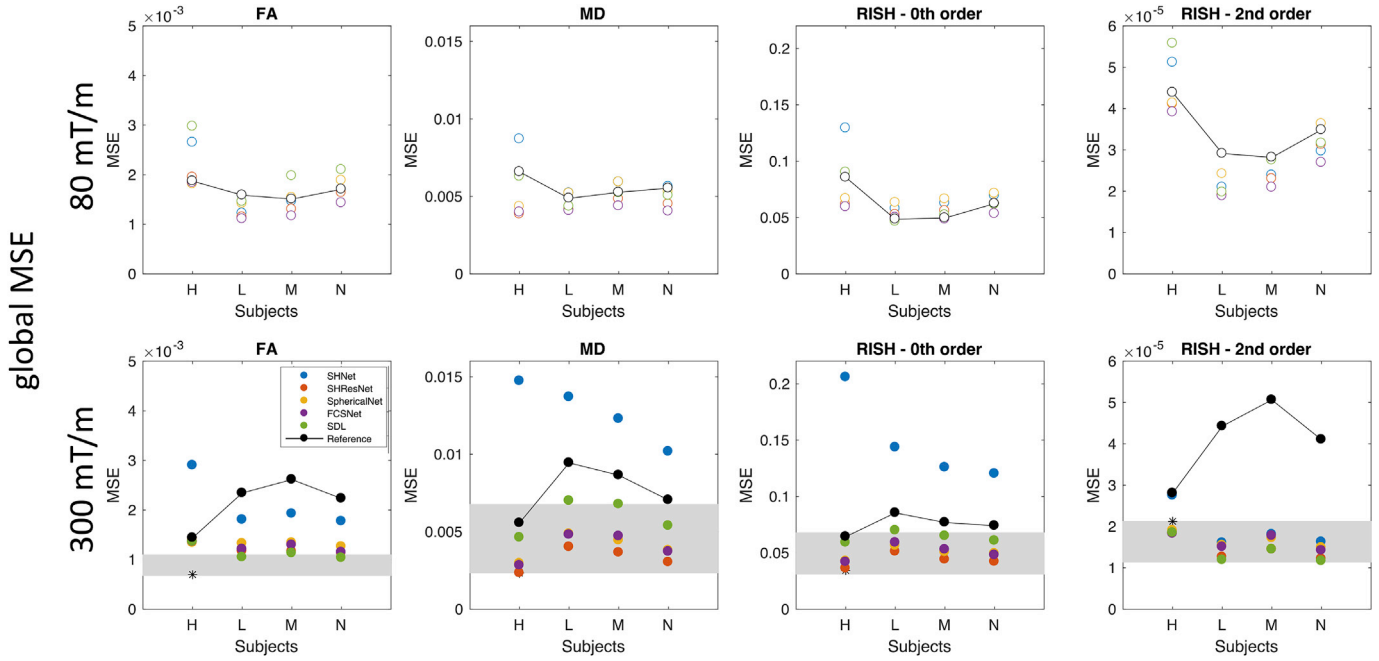

Fig. 6. Results of the matched resolution scanner-to-scanner mapping: Global MSE for FA, MD, R0, and R2 (columns); for the 4 different test subjects; for predictions of the 80 mT/m data (top, open circles) and the 300 mT/m data (bottom, closed circles). The different methods are represented with different colours SHNet (blue), SHResNet (red), SphericalNet (orange), FCSNet (purple), SDL (green), and reference (black). The reference results for different subjects are connected with lines as a visual aid to compare with the performance of the proposed algorithms. The range of scan-rescan MSEs across three subjects is shown by the gray rectangle. The scan-rescan result for the only overlapping subject "H" is indicated by the *.

gradient system and for three subjects, of which one subject overlapped with the evaluated subjects, i.e. "H". It can be observed that the best performing techniques approach the range of scan-rescan reproducibilities for all metrics, with an overall lower performance in terms of FA.

3.1.2. Regional evaluation

Fig. 7 shows the mean-squared error for 143 different white and gray matter ROIs and all evaluated algorithms. The median MSE per ROI and per evaluated algorithm across subject is shown. The Freesurfer ROIs are labeled according to Supplementary Table 3. Some ROIs have a consistently higher error, possibly due to residual image misalignment between different scanners. Indeed, the following ROIs scored in the WM/GM-specific 90th percentile for at least 3 algorithms, for at least one of the evaluated metrics: banks of the superior temporal sulcus, fusiform gyrus, temporal pole, transverse temporal gyrus, caudal anterior cingulate, insula, entorhinal, orbital part of inferior frontal gyrus, and frontal pole. However, no systematic trends in the major white matter bundles or the prefrontal cortex, where differences in susceptibility correction strategies could affect the result, were observed. Supplementary Table 3 marks systematically poor performing regions.

3.1.3. Local evaluation

Distributions of the localised MNE and MSE were computed across subjects per evaluated algorithm, Fig. 8 shows the median and width (95th percentile) of the MNE (top) and MSE (bottom) for 80 mT/m and 300 mT/m. This analysis was restricted to the white matter, and all ROIs that scored systematically poor (cf. Fig. 7) were excluded. Table 6 summarises the results.

The harmonisation algorithms generally have a median localised MNE closer to zero compared to the reference, and thus perform better than simple interpolation (Fig. 8, top). For each metric, at least one algorithm has a median localised MNE lower than 5%, with minimal errors less than 1% (Table 6). However, localised MNE of more than 15% are observed for all algorithms and evaluated metrics. The performance of the different harmonisation techniques varies widely (up to an order of magnitude) across the metrics and scanners, without one technique consistently outperforming all others. Indeed, 3 out of the 5 algorithms achieve the lowest median localised MNE error in at least one of the metrics. SHResNet, SphericalNet and FCSNet outperformed the reference for all metrics and both scanners, where FCSNet is the most consistent well-performing algorithm. Predictions for the 300 mT/m system are wider than for the 80 mT/m system (compare different bars in Fig. 8 top).

Fig. 8 bottom shows distributions of localised MSE evaluating both accuracy and precision, and most algorithms outperform the reference for both scanners. For 80 mT/m predictions, FCSNet outperformed all algorithms; for 300 mT/m predictions, SHResNet and SDL both outperformed the others in 2 of the 4 metrics.

3.2. Spatial- and angular resolution enhancement

3.2.1. Local evaluation

Fig. 9 shows the median and width of the localised MNE and MSE distributions, and the medians are also reported in Table 7. The performance for this task is poorer than the scanner-to-scanner mapping with errors being larger. The harmonisation algorithms do not always outperform the reference. For the 80 mT/m prediction, the reference interpolation even outperforms the harmonisation approaches for two out of four diffusion metrics, specifically the metrics related to anisotropy. For metrics MD and R0, SHResNet and SphericalNet consistently outperform the reference for both scanners. Overall, the 300 mT/m predictions have higher accuracy with a maximal median localised MNE of 7%. Again, a wide variability of all methods across the different metrics and scanners could be observed without one method outperforming the others.

4. Discussion

With the increasing prevalence of MRI research systems capable of collecting diffusion MRI data, comes the potential of combining data sets

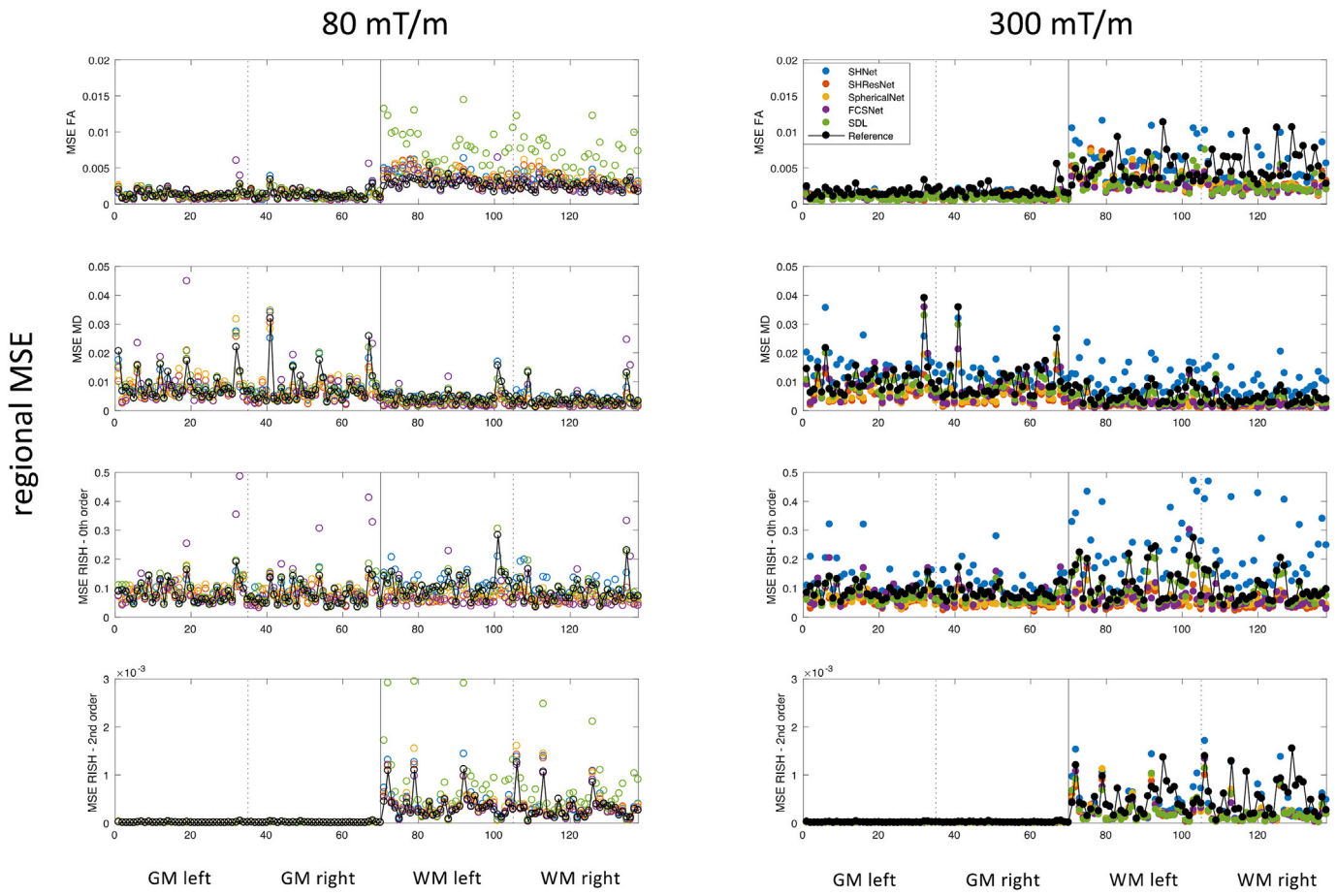

Fig. 7. Results of the scanner-to-scanner mapping: Regional MSE for FA, MD, R0, and R2 (rows); median MSE across subjects per ROI per algorithm; for predictions of the 80 mT/m data (left, open circles) and the 300 mT/m data (right, closed circles). The different methods are represented with different colours SHNet (blue), SHResNet (red), SphericalNet (orange), FCSNet (purple), SDL (green), reference (black). Regions 1 to 35 and 36 to 71 are left and right GM regions respectively, regions 72 to 107 and 108 to 143 are left and right WM regions respectively (see Supplementary Table 3).

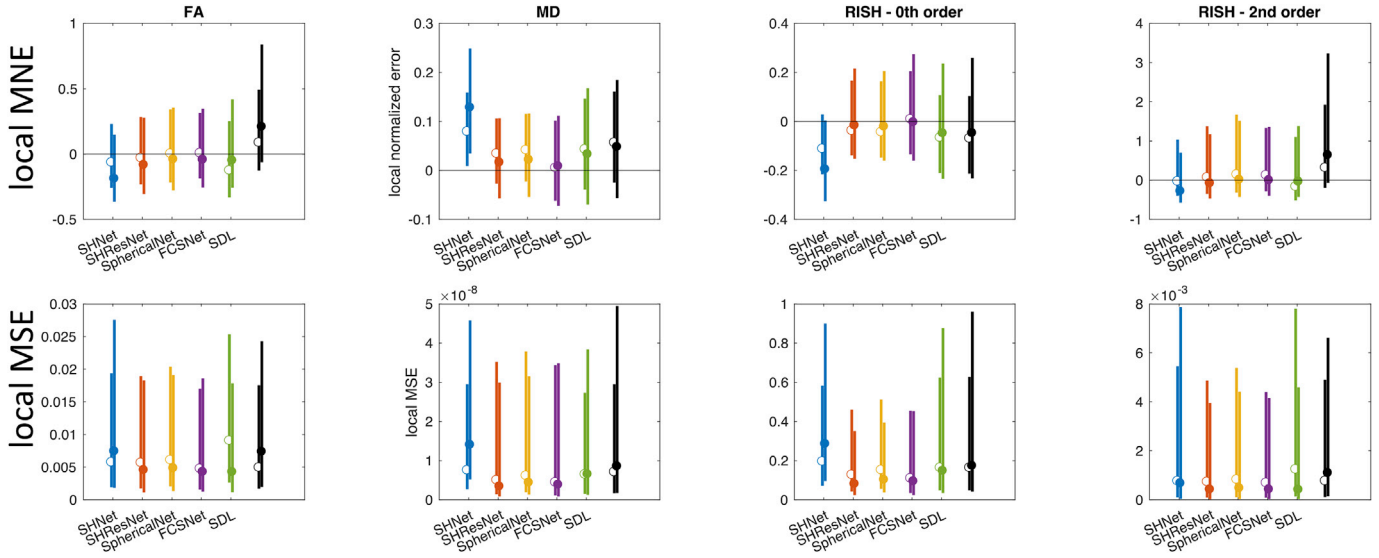

Fig. 8. Results of the scanner-to-scanner mapping: Local MSE and MNE for FA, MD, R0, and R2 (columns), across WM (excluding problematic ROIs) and all subjects. The different methods are represented with different colours: SHNet (blue), SHResNet (red), SphericalNet (orange), FCSNet (purple), SDL (green), and reference (black). (a) Local MNE distributions for the 80 mT/m predictions (top) and 300 mT/m predictions (bottom). (b) The bars show the 95th percentile of the MNE and MSE distributions for the 80 mT/m predictions (median indicated by open circles) and the 300 mT/m predictions (median indicated by closed circles).

Table 6
Results of the scanner-to-scanner mapping: median localised MNE and median localised MSE for each algorithm and metric, across subjects. Algorithms performing better than the reference (trilinear and SH interpolation) are highlighted in gray, the best performing algorithm for each metric is highlighted in red.

INME(%)	80 mT/m				300 mT/m			
	FA	MD	R0	R2	FA	MD	R0	R2
SHNet	-6.49	7.90	-11.23	-3.20	-18.93	12.83	-19.56	-28.29
SHResNet	-3.00	3.39	-3.81	6.82	-8.41	1.60	-1.68	-8.12
SphericalNet	0.27	4.15	-4.38	14.22	-4.12	2.14	-2.12	0.82
FCSNet	0.65	0.48	0.90	12.56	-4.35	0.84	-0.37	-0.37
SDL	-12.58	4.28	-6.66	-16.67	-5.02	3.28	-4.86	-3.83
Reference	8.68	5.65	-6.94	31.39	20.85	4.77	-4.76	63.86

IMSE (x1000)	80 mT/m				300 mT/m			
	FA	MD	R0	R2	FA	MD	R0	R2
SHNet	5.72	7.50E-06	195.01	0.76	7.41	14.0E-06	284.63	0.668
SHResNet	5.64	4.99E-06	126.82	0.73	4.53	3.38E-06	80.04	0.414
SphericalNet	6.06	6.12E-06	151.44	0.82	4.81	4.35E-06	101.85	0.471
FCSNet	4.75	4.42E-06	108.81	0.69	4.24	3.82E-06	94.08	0.419
SDL	9.03	6.45E-06	163.07	1.24	4.23	6.49E-06	147.43	0.406
Reference	4.92	7.04E-06	163.04	0.76	7.35	8.51E-06	172.80	1.090

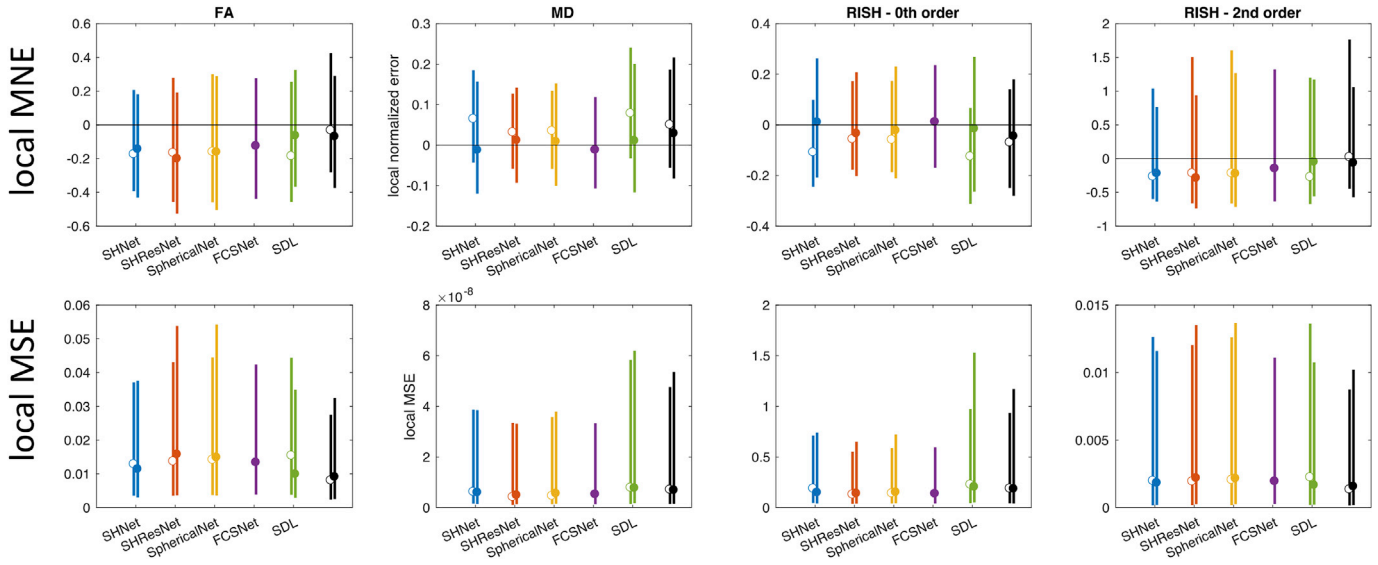

Fig. 9. Results of the spatial and angular resolution enhancement: Local MSE and MNE for FA, MD, R0, and R2 (columns), across WM (excluding problematic ROIs) and all subjects. The different methods are represented with different colours: SHNet (blue), SHResNet (red), SphericalNet (orange), FCSNet (purple), SDL (green), and reference (black). The bars show the 95th percentile of the MNE and MSE distributions for the 80 mT/m predictions (median indicated by open circles) and the 300 mT/m predictions (median indicated by closed circles).

Table 7
Results of the spatial and angular resolution enhancement: median localised MNE and median localised MSE for each algorithm and metric, across subjects. Algorithms performing better than the reference are highlighted in gray, the best performing algorithm for each metric is highlighted in red.

INME(%)	80 mT/m				300 mT/m			
	FA	MD	R0	R2	FA	MD	R0	R2
SHNet	-17.50	6.46	-10.90	-26.78	-14.44	-1.27	1.04	-22.16
SHResNet	-16.62	3.15	-5.67	-21.90	-20.07	1.15	-3.41	-28.97
SphericalNet	-16.08	3.48	-5.90	-21.78	-16.15	0.83	-2.30	-22.55
FCSNet	N/A	N/A	N/A	N/A	-12.55	-1.21	1.14	-14.8
SDL	-18.58	7.82	-12.54	-27.50	-6.47	1.04	-1.55	-4.91
Reference	-3.33	5.00	-7.00	2.08	-6.99	2.87	-4.53	-6.55

IMSE (x1000)	80 mT/m				300 mT/m			
	FA	MD	R0	R2	FA	MD	R0	R2
SHNet	12.87	6.21E-06	187.94	1.98	11.40	5.97E-06	148.35	1.83
SHResNet	13.70	4.14E-06	129.28	1.94	15.76	4.88E-06	139.05	2.19
SphericalNet	14.18	4.62E-06	140.28	2.05	14.87	5.55E-06	152.36	2.16
FCSNet	N/A	N/A	N/A	N/A	13.37	5.22E-09	136.69	1.95
SDL	15.43	7.86E-06	226.89	2.24	9.90	7.72E-06	203.66	1.66
Reference	8.05	7.10E-06	188.28	1.34	9.09	6.88E-06	184.37	1.56

of much larger size than could ever be collected at one centre alone. Several studies have reported a variability between diffusion measurements acquired at different scanners and sites (Chen et al., 2014; Mirzaalian et al., 2017, 2016; Vollmar et al., 2010), even with comparable protocols. When protocols differ substantially because of updated hardware and software, more 'historical' data can potentially be enhanced by learning features from 'state-of-the-art' data (Alexander et al., 2017).

As a result, the interest in developing methods to establish a mapping between different scanners and protocols is continuously growing (Jahanshad et al., 2013; Kochunov et al., 2014; Venkatraman et al., 2015; Jenkins et al., 2016; Pohl et al., 2016; Fortin et al., 2017; Mirzaalian et al., 2016). Harmonisation on groups of different subjects scanned on different scanners can be performed by finding spatial correspondence between subjects by registration to an atlas, and relies on the assumption that the diffusion measurements between matched groups (in age, gender, etc.) are statistically different only due to scanner-differences. In a group of travelling subjects as presented here, the confounding factor of inter-subject differences is removed and spatial correspondence can be obtained more directly. While a travelling control group does not preclude group-level harmonisation strategies, it should allow scanner-specific effects to be captured with fewer subjects. In this work, we have presented a benchmarking database of acquisitions of the same healthy controls scanned on different scanners with different maximum gradient strengths and protocols.

4.1. Data

Scanning a 'travelling head' on different scanners should ideally be performed in quick succession to avoid intermediate software updates and age-related effects. In this work, the average time between acquisitions on scanners a) and b), and a) and c) was 21 and 22 months, respectively. Whereas the time between scans b) and c) was very short, the time difference with scan a) was longer. Age-related changes during this period might be present but are assumed to be small compared to the source of variance introduced by cross-scanner and cross-protocol differences. By including adult subjects with an average age of 25.7 years we have strived to minimise such confounds; previous studies have shown that age-related FA and MD changes in several white matter structures reach a plateau around the age of 25 (Lebel et al., 2008). Similarly, variabilities that might occur when scanning subjects at different time points during a day (Thomas et al., 2018) are assumed to be small compared to cross-scanner variability in this study. None of the scanners had software upgrades during the course of the study.

Scan-rescan experiments in quick succession can provide information on the inherent measurement variability of a particular scanner and sequence, and as such give an estimate of the lower bound of harmonisation performance. We performed a preliminary analysis on rescans of the $b = 1200\,s/mm^2$ data for 3 subjects on the 300 mT/m system, of which one subject overlapped with the evaluated subjects. The global MSE (Fig. 6, bottom) shows that the best performing techniques approach the range of scan-rescan reproducibilities for all metrics, with an overall lower performance in terms of FA. Median localised MNE (not shown) varied from 2 to 5% for FA in agreement with the literature (Kochunov et al., 2014). We have adopted the same normalization procedure of registering the data to the 80 mT/m space (both the scan- and rescan). Better overlap of inter-scanner data could be achieved than for intra-scanner data, which might have partially contributed to lower median MSE values.

In addition to travelling heads, physical phantoms can be used to detect scanner-specific variabilities and changes, and to correct for such variabilities with harmonisation approaches. While such phantoms do not suffer from age-related or time-of-day effects, they are incapable of fully capturing the complexity of biological tissue and regional differences associated with this complexity. Furthermore, it can be non-trivial to translate the differences observed in physical phantoms to in-vivo acquisitions. Nevertheless, learning a scanner-to-scanner mapping from a travelling phantom would be an interesting alternative challenge, and ideally both in-vivo travelling head- and physical phantom acquisitions could be combined to assess variabilities and evaluate harmonisation approaches.

The acquisition parameters of the matching resolution cross-scanner (ST) protocols were harmonised as closely as possible in terms of b-value, TE, TR, spatial resolution, and angular resolution. However, slight variations between the ST protocols on different scanners and vendors remained. While this could introduce additional variability in the measurements, it also mimics common more subtle variabilities between scanner sites. This allows us to test whether harmonisation approaches are robust to such changes.

While the database was here used as a testbed for harmonisation algorithms, it could contribute to answering alternative questions as well. Recently, part of the database has been utilised to investigate the dependency of Meyer's loop tractography on imaging protocol and hardware (Chamberland et al., 2018).

4.2. Preprocessing

Minimally preprocessed data were made available to the entrants of the challenge, but the raw (unprocessed) data will be made available upon request. The data preprocessing pipeline can have an effect on the degree to which the datasets are comparable prior to data harmonisation (Jenkins et al., 2016), as differences in preprocessing can induce differences in data across acquisitions that are hard to harmonise *a posteriori*. For example, different methods were used to correct for susceptibility distortions between the 40 mT/m system and the other systems, because a reversed phase encoding b0 image was not available for the former. In the current study, we did not find that regions affected by susceptibility distortions (e.g. frontal regions) performed systematically poorer than other regions, but such differences will likely have an effect when performing multi-centre studies. Investigating this effect is subject to future work.

While the preprocessing pipeline included the most commonly performed steps such as motion correction and eddy current- and susceptibility distortion correction, the importance of correcting for other artifacts has been stressed in various works; e.g. Gibbs ringing (Kellner et al., 2016; Perrone et al., 2015; Veraart et al., 2016a), signal drift (Vos et al., 2016), and others (Andersson, 2014; Le Bihan et al., 2006; Pierpaoli, 2010; Tax et al., 2016). Manual inspection of the data did not reveal any gross artifacts such as slice intensity dropouts, but artifacts could have a less visible impact on signal intensities and as such additional preprocessing steps could be included in the preprocessing pipeline.

The use of different software tools for the different preprocessing steps resulted in multiple re-samplings and interpolations of the data. Ideally, the different warps (of motion/eddy current distortion correction, gradient nonlinearity distortion correction, and registration to a common space) should be concatenated and performed within a single interpolation step. However, to the best of our knowledge, there is currently no consensus on the order of performing motion/eddy current distortion correction and gradient nonlinearity distortion correction (Rudrapatna et al., 2018), and the 'best' practice likely depends on the amount of subject motion.

Gradient nonlinearity in the 300 mT/m system not only causes geometrical image distortions, but also spatially varying b-vectors and b-values (Bammer et al., 2003). To simplify the tasks, the variance in diffusion weighting was not taken into account in the current comparison, and might be a possible explanation for the greater variance in the 300 mT/m predictions. This information could potentially improve harmonisation with the 300 mT/m system and will be made available.

For the registration to a common space, we have qualitatively compared different toolboxes, degrees of freedom (linear vs affine), and input images (FA, B0, mean of the DWI), and observed that the quality of registration varied with the input image and degrees of freedom. Here, we

used the mean of the DWI images, but other choices are possible, such as b0 images, FA, or a multi-contrast approach. Rigid (translation and rotation) registration to a common space for each subject was generally not sufficient to achieve good alignment. This indicates that residual distortions remain that were not corrected during preprocessing. Nonlinear registration was used for the 40 mT/m data to simultaneously correct for susceptibility distortions in the phase encoding direction, while affine registration was used for all the other data. Full nonlinear registration is envisioned to give better overlap, but the interaction between local deformations and the orientational information present in the DWIs adds a layer of complexity. Therefore, for this work, we decided to stick to affine transformations, but addressing the remaining distortions between scanners after preprocessing is an important issue general to harmonisation that should be addressed in future work. In Supplementary Material 4, we show a qualitative comparison of the registrations for one subject, and we report the decrease of mean squared error that could be achieved with full nonlinear registration of the test subjects, compared to the registration performed in the current evaluation. The MSE were all lower than 0.016, and the 300 mT/m ST registration showed the lowest improvement in MSE with full nonlinear registration, followed by 80 mT/m SA, 300 mT/m SA, and 40 mT/m ST. We also report the mean of the Jacobian determinant, reflecting how much local deformation was necessary.

4.3. Evaluation procedure

The evaluation in this study was performed on a global, regional, and local level, and different diffusion metrics were derived. The results suggest that the relative performance of algorithms strongly depends on the metric evaluated, and that the harmonisation can thus be tuned towards the metric of interest. The current work focused on the presentation of the harmonisation-benchmark database and a first evaluation of harmonisation algorithms on this database, where the evaluation was specifically targeted at metrics that are most widely used in clinical studies (e.g. DTI features from single-shell data). This can be extended in future work in multiple ways. Evaluation of higher-order metrics, e.g. RISH metrics derived from the 4^{th} and 6^{th} spherical harmonic order, could provide further insight into the performance of the signal harmonisation, but the precision of their estimates can be lower. For higher order metrics, it would be interesting to investigate the impact of the number of gradient directions when the resolution is higher and the SNR lower, as is the case in the resolution enhancement task. In addition, as multi-shell dMRI data is becoming more readily available, harmonisation algorithms should be extended to accommodate data acquired with multiple b-values. The presented database includes dMRI data with multiple b-values, and therefore allows an evaluation of such algorithms in future work. Finally, the diffusion signals could be compared directly. However, the actual diffusion measurements include B1 transmit-, amplifier-, and receive effects, and their multiplicative scaling is arbitrary. An additional prediction of this scale for each scan would be necessary and therefore adds another layer of complexity. In this first evaluation we have therefore opted for the evaluation of summary metrics that are independent of this scale, but the evaluation of individual predicted diffusion signals can be performed in future work.

In this work, the aim was to harmonise the DWIs directly in native space (as for example also done in Mirzaalian et al. (2016)), as opposed to harmonising feature maps such as DTI-derived metrics (Fortin et al., 2017; Jahanshad et al., 2013; Kochunov et al., 2014; Pohl et al., 2016; Venkatraman et al., 2015). The direct harmonisation of DWIs is beneficial in that any metric can later be compared between groups, and it potentially allows to better capture certain complex variations as opposed to harmonising particular features alone. Performing the mapping in atlas space can be beneficial if the data is analysed in this common space to avoid the additional back-projection step to native space (e.g. in voxel-based analysis), but harmonisation in native space allows the use of approaches that analyse the data without registration to an atlas, e.g. tract-based approaches where tractography is performed and mean-tract values are compared across subjects. In this work, we evaluate predictions of the data at the reference site and compare to the acquired ground truth in native space. Evaluating algorithms by their ability to remove statistical group differences would be possible, but would likely benefit from larger cohorts. The challenge setup is therefore in part motivated by the sample size.

MNE and MSE were computed to assess accuracy and precision. In addition to comparing low-level error metrics and reduce this down to a score for the overall performance of a harmonisation approach, an alternative evaluation can be targeted to a particular task; for example, the harmonisation of values along selected tracts reconstructed with tractography or the ability to discriminate patients and controls. While this would give a clearer picture of harmonisation performance for the task at hand, it remains unclear how these results extend among different tasks.

Misalignments between scanners both in the training data and test data can influence the results. For this reason, a registration step was included in the preprocessing pipeline so that misregistration would affect all harmonisation algorithms to the same extent. While misalignments can vary in a non-systematic way owing to inter-individual differences, systematically bad performing regions may have suffered from this to a larger extent. None of the entrants reported additional registration steps, so it can be expected that misregistration would affect all harmonisation algorithms to a similar degree. It is envisioned that optimised registration will not only improve the learned mapping between scanners, but may also reduce the errors observed in the evaluation stage.

4.4. Implications and recommendations

From the results, some broad trends become apparent. For the scanner-to-scanner mapping, the best performing algorithms give a good consistency across subjects (in terms of global MSE, Fig. 6), but less consistency across regions (Fig. 7). Fig. 7 shows differences in regional error scores for all algorithms between WM and GM: metrics that describe anisotropy (i.e. FA and R2) have larger MSE in WM, whereas MD and R0 have more comparable MSE in WM vs GM. For most metrics and harmonisation algorithms, median localised MNE of $<5\%$ could be obtained for both scanners (with values of $<1\%$ for the best performing algorithms), however, also localised MNE of $>15\%$ were observed.

For the angular- and spatial-resolution enhancement, median localised MNE of $<5\%$ and $<1\%$ could be obtained for the 80 mT/m and 300 mT/m scanner respectively, but only for isotropic measures (MD and R0). A possible explanation is that local variations in MD and R0 are smaller (e.g. at WM/GM interfaces) such that spatial misalignment problems could be less pronounced. Machine learning techniques probably outperform linear interpolation because they are able to correct global offsets. In contrast, FA and R2 are anisotropy measures that have clear 'edges' at WM/GM interfaces, hence co-registration issues can be amplified. FCSNet is most consistently well-performing, and its use of a larger local neighbourhood could be beneficial to ameliorate such issues. The good performance of the reference interpolation compared to machine learning-based methods might indicate that some blurring might improve the results, but too much blurring is detrimental in most cases (Supplementary Fig. 1).

It is important to consider these numbers in the light of the magnitudes of the effects-of-interest in group studies, which are often smaller than the 15% mentioned above. While harmonisation approaches should remove differences across sites, inter-individual differences and differences between the healthy control and disease population should be preserved after harmonisation. We see our database as a first significant step towards a general benchmarking of harmonisation methods in healthy controls. A breadth of different microstructural architectural paradigms can already be seen across the young healthy brain; from isotropic CSF, to isotropic gray matter, to white matter with multiple crossing fibres, to white matter with single fibres and different degrees of anisotropy, as well as different degrees of partial voluming with isotropic configurations. Depending on the degree of microstructural changes, this

may or may not fall within the range of microstructural architectural paradigms that can be predicted from healthy controls. Nevertheless, it would be reasonable to expect that if harmonisation algorithms have been reported to work well for the dataset present here, that they will likely work well on training databases that include a wider age range and pathological cases. This could also potentially be assessed by synthetic experiments and data augmentation, as e.g. in Mirzaalian et al. (2016).

To assess the impact on real-world analysis tasks, such as group- or longitudinal studies where large samples are necessary to detect small effects, power analyses could be performed. For example, one could work out the reduction in the total number of participants needed across the scanners to see a given effect size. Alternatively, for a fixed number of participants, one could compute what effect size could be detected. The exact computation depends on the choice of effect size (Lakens, 2013). Scan-rescan errors can serve as a good estimate of the inherent variability and lower bound performance of all harmonisation algorithms, and therefore of the smallest changes that can realistically be observed. The current results suggest that in scanner-to-scanner mapping, harmonisation approaches can reduce the variability for the metrics investigated, with the best performing algorithms approaching the range of scan-rescan reproducibilities.

Based on the exercise of data harmonisation as performed in this study, we highlight a few recommendations for data harmonisation in future multi-centre studies, covering different stages of the analysis pipeline:

1) Variability between acquisitions on different scanners should be reduced by matching of protocols as much as possible. This can be a challenge in itself as different platforms do not always allow perfect matching. One should consider main parameters such as spatial resolution, TE, number and distribution of gradient directions, and b-value; but also other parameters such as parallel imaging, multiband, partial Fourier, and reconstruction settings. It is envisioned that better matching will lead to better *a priori* harmonisation, but a reduction in measurement variability can still be achieved in the case of non-perfect matching, as illustrated in this study. The question which parameters have the largest effect is challenging to answer and can be the subject of future studies.
2) Because of the effect sizes typically found in diffusion MRI studies, a travelling head and/or travelling phantom study should ideally be performed on all the scanners involved in the study, as well as a characterisation of scan-rescan variability. This requires subjects to be scanned and re-scanned across different sites and between protocols, but potentially also before and after every software update. Training datasets on the order of 10 subjects, as used in this work, could be a logistic and financial burden on a study if the sample size is small. However, it could be argued that if a research question requires a seriously large sample size that demands a multi-site study, the proportional cost and administrative burden of scanning a few subjects on both sites is manageable and could be justified by the advantage of reliable harmonisation and increased statistical power. Unless the patient group is exceedingly rare and a multi-site study is required to achieve even a small sample size, studies with small sample sizes may as well be performed at a single site. Future work should assess this trade-off, and compare to a group-based approach where different subjects scanned on different scanners are matched and used for harmonisation.
3) Accurate characterisation and detection of artifacts on each scanner is important to improve *a priori* harmonisation (that is, prior to applying harmonisation algorithms as described in this study). To this end, acquiring additional data for improved artefact correction is recommended. One can think of acquiring reversed-phase encoding images or field maps for the correction of susceptibility distortions (Andersson et al., 2003; Irfanoglu et al., 2015), noise maps for the characterisation of noise distributions (Froeling et al., 2016) and denoising (e.g. (St-Jean et al., 2016; Veraart et al., 2016c)), and additional gradient directions to improve outlier detection and reduce the effect of outlier rejection (Chang et al., 2005; Collier et al., 2015; Mangin et al., 2002; Sairanen et al., 2018; Tax et al., 2015). Not appropriately correcting for such artifacts can result in increased variability and can impose extra challenges to harmonisation algorithms.
4) Likewise, differences in preprocessing and model estimation strategies between different scanners can introduce additional variability, and such pipelines should be matched whenever possible.
5) Point-to-point mapping between different subjects and different scanners by means of a registration procedure deserves attention; residual distortions after preprocessing can be observed.
6) In the case of a clear hypothesis of the region or tract involved in the phenomenon under investigation one could perform a more in-depth investigation of measurement variability than global error measures.

4.5. Obtaining the data

The database is available to the public, to encourage further development of harmonisation approaches and to further evaluate e.g. the use of machine learning vs alternative methods, using spatial context vs not using spatial context, and employing data augmentation techniques. Information on how to obtain the data can be found on the following webpage: https://www.cardiff.ac.uk/cardiff-university-brain-research-imaging-centre/research/projects/cross-scanner-and-cross-protocol-diffusion-MRI-data-harmonisation.

5. Conclusion

In conclusion, cross-scanner and cross-protocol measurement variability challenges multi-centre studies. In this study, we have presented a benchmarking database to test and evaluate harmonisation approaches for cross-scanner and cross-protocol mapping. The harmonisation approaches proposed significantly reduce variability in multi-vendor diffusion scans with comparable protocols, but challenges in spatial- and angular resolution enhancement of features that characterise anisotropy remain. Before widespread deployment of harmonisation schemes for large multi-site studies, it would be important to perform more rigorous testing, e.g. in other training sets. In future work, the benchmarking database will be utilised to evaluate harmonisation approaches for multi-shell diffusion MRI data and the influence of preprocessing on the performance of harmonisation.

Acknowledgements

CMWT is supported by a Rubicon grant (680-50-1527) from the Netherlands Organisation for Scientific Research and Wellcome Trust grant (096646/Z/11/Z). This project has received funding under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634541 and from Engineering and Physical Sciences Research Council (EPSRC EP/R006032/1, M020533/1), funding FG. SSJ is supported by the Fonds de Recherche du Québec - Nature et Technologies (Dossier 192865). AL and SSJ are supported by VIDI Grant 639.072.411 from the Netherlands Organisation for Scientific Research. RT acknowledges funding from Microsoft Research. DCA and AG acknowledge funding from EPSRC grants N018702 M020533 L022680. DEJL and DKJ were supported by MRC grant MR/K004360/1. Scan costs were supported by the National Centre for Mental Health (NCMH) with funds from Health and Care Support Wales and by the Wellcome Trust. JV is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO; grant number 12S1615N). LN is supported in part by NIH grants R21MH115280 and R21MH116352.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.01.077.

References

Alexander, D.C., Zikic, D., Ghosh, A., Tanno, R., Wottschel, V., Zhang, J., Kaden, E., Dyrby, T.B., Sotiropoulos, S.N., Zhang, H., Criminisi, A., 2017. Image quality transfer and applications in diffusion MRI. Neuroimage 152, 283–298. https://doi.org/10.1016/J.NEUROIMAGE.2017.02.089.

Andersson, J.L.R., 2014. Chapter 4 – geometric distortions in diffusion MRI. In: Diffusion MRI, pp. 63–85. https://doi.org/10.1016/B978-0-12-396460-1.00004-4.

Andersson, J.L.R., Jenkinson, M., Smith, S., 2007. Non-linear Registration Aka Spatial Normalisation. Oxford.

Andersson, J.L.R., Skare, S., Ashburner, J., 2003. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage 20, 870–888. https://doi.org/10.1016/S1053-8119(03)00336-7.

Andersson, J.L.R., Sotiropoulos, S.N., 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 125, 1063–1078. https://doi.org/10.1016/j.neuroimage.2015.10.019.

Bammer, R., Markl, M., Barnett, A., Acar, B., Alley, M.T., Pelc, N.J., Glover, G.H., Moseley, M.E., 2003. Analysis and generalized correction of the effect of spatial gradient field distortions in diffusion-weighted imaging. Magn. Reson. Med. 50, 560–569. https://doi.org/10.1002/mrm.10545.

Bischoff-Grethe, A., Ozyurt, I.B., Busa, E., Quinn, B.T., Fennema-Notestine, C., Clark, C.P., Morris, S., Bondi, M.W., Jernigan, T.L., Dale, A.M., Brown, G.G., Fischl, B., 2007. A technique for the deidentification of structural brain MR images. Hum. Brain Mapp. 28, 892–903. https://doi.org/10.1002/hbm.20312.

Chamberland, M., Tax, C.M.W., Jones, D.K., 2018. Meyer's loop tractography for image-guided surgery depends on imaging protocol and hardware. NeuroImag. Clin. https://doi.org/10.1016/J.NICL.2018.08.021.

Chang, L.-C., Jones, D.K., Pierpaoli, C., 2005. RESTORE: robust estimation of tensors by outlier rejection. Magn. Reson. Med. 53, 1088–1095. https://doi.org/10.1002/mrm.20426.

Chen, J., Liu, J., Calhoun, V.D., Arias-Vasquez, A., Zwiers, M.P., Gupta, C.N., Franke, B., Turner, J.A., 2014. Exploration of scanning effects in multi-site structural MRI studies. J. Neurosci. Methods 230, 37–50. https://doi.org/10.1016/J.JNEUMETH.2014.04.023.

Collier, Q., Veraart, J., Jeurissen, B., den Dekker, A.J., Sijbers, J., 2015. Iterative reweighted linear least squares for accurate, fast, and robust estimation of diffusion magnetic resonance parameters. Magn. Reson. Med. 73, 2174–2184. https://doi.org/10.1002/mrm.25351.

de Lange, E.E., Mugler, J.P., Bertolina, J.A., Gay, S.B., Janus, C.L., Brookeman, J.R., 1991. Magnetization Prepared RApid Gradient-Echo (MP-RAGE) MR imaging of the liver: comparison with spin-echo imaging. Magn. Reson. Imaging 9, 469–476. https://doi.org/10.1016/0730-725X(91)90031-G.

Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R., 2007. Regularized, fast, and robust analytical Q-ball imaging. Magn. Reson. Med. 58, 497–510. https://doi.org/10.1002/mrm.21277.

Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Glasser, M.F., Miller, K.L., Ugurbil, K., Yacoub, E., 2010. Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. PLoS One 5, e15710. https://doi.org/10.1371/journal.pone.0015710.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X.

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, 149–170. https://doi.org/10.1016/J.NEUROIMAGE.2017.08.047.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Software 33, 1–22.

Froeling, M., Tax, C.M., Vos, S.B., Luijten, P.R., Leemans, A., 2016. "MASSIVE" brain dataset: multiple acquisitions for standardization of structural imaging validation and evaluation. Magn. Reson. Med. https://doi.org/10.1002/mrm.26259.

Gallichan, D., Scholz, J., Bartsch, A., Behrens, T.E., Robson, M.D., Miller, K.L., 2009. Addressing a systematic vibration artifact in diffusion-weighted MRI. Hum. Brain Mapp. 31 https://doi.org/10.1002/hbm.20856. NA-NA.

Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., van der Walt, S., Descoteaux, M., Nimmo-Smith, I., 2014. Dipy, a library for the analysis of diffusion MRI data. Front. Neuroinf. 8, 8. https://doi.org/10.3389/fninf.2014.00008.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124. https://doi.org/10.1016/J.NEUROIMAGE.2013.04.127.

Golkov, V., Dosovitskiy, A., Sämann, P., Sperl, J.I., Sprenger, T., Czisch, M., Menzel, M.I., Gómez, P.A., Haase, A., Brox, T., Cremers, D., 2015. Q-Space Deep Learning for Twelve-fold Shorter and Model-free Diffusion MRI Scans. Springer, Cham, pp. 37–44. https://doi.org/10.1007/978-3-319-24553-9_5.

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., Peet, A.C., Clark, C.A., 2015. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. NMR Biomed. 28, 468–485. https://doi.org/10.1002/nbm.3269.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Irfanoglu, M.O., Modi, P., Nayak, A., Hutchinson, E.B., Sarlls, J., Pierpaoli, C., 2015. DR-BUDDI (Diffeomorphic Registration for Blip-Up blip-Down Diffusion Imaging) method for correcting echo planar imaging distortions. Neuroimage 106, 284–299. https://doi.org/10.1016/j.neuroimage.2014.11.042.

Jahanshad, N., Kochunov, P.V., Sprooten, E., Mandl, R.C., Nichols, T.E., Almasy, L., Blangero, J., Brouwer, R.M., Curran, J.E., de Zubicaray, G.I., Duggirala, R., Fox, P.T., Hong, L.E., Landman, B.A., Martin, N.G., McMahon, K.L., Medland, S.E., Mitchell, B.D., Olvera, R.L., Peterson, C.P., Starr, J.M., Sussmann, J.E., Toga, A.W., Wardlaw, J.M., Wright, M.J., Hulshoff Pol, H.E., Bastin, M.E., McIntosh, A.M., Deary, I.J., Thompson, P.M., Glahn, D.C., 2013. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA–DTI working group. Neuroimage 81, 455–469. https://doi.org/10.1016/J.NEUROIMAGE.2013.04.061.

Jenkins, J., Chang, L.-C., Hutchinson, E., Irfanoglu, M.O., Pierpaoli, C., 2016. Harmonization of methods to facilitate reproducibility in medical data processing: applications to diffusion tensor magnetic resonance imaging. In: 2016 IEEE International Conference on Big Data (Big Data). IEEE, pp. 3992–3994. https://doi.org/10.1109/BigData.2016.7841086.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62, 782–790. https://doi.org/10.1016/J.NEUROIMAGE.2011.09.015.

Johansen-Berg, H., Behrens, T.E.J., 2009. Diffusion MRI: from Quantitative Measurement to In-Vivo Neuroanatomy, Diffusion MRI. https://doi.org/10.1016/B978-0-12-374709-9.00002-X.

Jones, D.K., 2010a. Diffusion MRI Theory, Methods, and Applications. Oxford University Press.

Jones, D.K., 2010b. Precision and accuracy in diffusion tensor magnetic resonance imaging. Top. Magn. Reson. Imag. 21, 87–99. https://doi.org/10.1097/rmr.0b013e31821e56ac.

Jones, D.K., Alexander, D.C., Bowtell, R., Cercignani, M., Dell'Acqua, F., McHugh, D.J., Miller, K.L., Palombo, M., Parker, G.J.M., Rudrapatna, U.S., Tax, C.M.W., 2018. Microstructural imaging of the human brain with a 'super-scanner': 10 key advantages of ultra-strong gradients for diffusion MRI. Neuroimage. https://doi.org/10.1016/J.NEUROIMAGE.2018.05.047.

Jones, D.K., Leemans, A., 2011. Diffusion tensor imaging. Methods Mol. Biol. 711, 127–144. https://doi.org/10.1097/01.chi.0000246064.93200.e8.

Karayumak, Suheyla Cetin, Bouix, Sylvain, Ning, Lipeng, James, Anthony, Crow, Tim, Shenton, Martha, Kubicki, Marek, Rathi, Yogesh, 2019. Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. NeuroImage 184, 180–200. ISSN 1053-8119.

Kellner, E., Dhital, B., Kiselev, V.G., Reisert, M., 2016. Gibbs-ringing artifact removal based on local subvoxel-shifts. Magn. Reson. Med. 76, 1574–1581. https://doi.org/10.1002/mrm.26054.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. In: Proc. 3rd Int. Conf. Learn. Represent..

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. 29, 196–205. https://doi.org/10.1109/TMI.2009.2035616.

Kochunov, P., Jahanshad, N., Sprooten, E., Nichols, T.E., Mandl, R.C., Almasy, L., Booth, T., Brouwer, R.M., Curran, J.E., de Zubicaray, G.I., Dimitrova, R., Duggirala, R., Fox, P.T., Elliot Hong, L., Landman, B.A., Lemaitre, H., Lopez, L.M., Martin, N.G., McMahon, K.L., Mitchell, B.D., Olvera, R.L., Peterson, C.P., Starr, J.M., Sussmann, J.E., Toga, A.W., Wardlaw, J.M., Wright, M.J., Wright, S.N., Bastin, M.E., McIntosh, A.M., Boomsma, D.I., Kahn, R.S., den Braber, A., de Geus, E.J.C., Deary, I.J., Hulshoff Pol, H.E., Williamson, D.E., Blangero, J., van 't Ent, D., Thompson, P.M., Glahn, D.C., 2014. Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling. Neuroimage 95, 136–150. https://doi.org/10.1016/J.NEUROIMAGE.2014.03.033.

Koppers, S., Bloy, L., Berman, J.I., Tax, C.M.W., Edgar, J.C., Merhof, D., 2018. Spherical harmonic residual network for diffusion signal harmonization. In: Computational Diffusion MRI. Springer.

Koppers, S., Haarburger, C., Merhof, D., 2017. Diffusion MRI Signal Augmentation: from Single Shell to Multi Shell with Deep Learning. Springer, Cham, pp. 61–70. https://doi.org/10.1007/978-3-319-54130-3_5.

Koppers, S., Merhof, D., 2018. DELIMIT PyTorch - an Extension for Deep Learning in Diffusion Imaging arXiv:1808.01517.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863. https://doi.org/10.3389/fpsyg.2013.00863.

Larkman, D.J., Hajnal, J.V., Herlihy, A.H., Coutts, G.A., Young, I.R., Ehnholm, G., 2001. Use of multicoil arrays for separation of signal from multiple slices simultaneously excited. J. Magn. Reson. Imag. 13, 313–317. https://doi.org/10.1002/1522-2586(200102)13:2<313::AID-JMRI1045>3.0.CO;2-W.

Le Bihan, D., Poupon, C., Amadon, A., Lethimonnier, F., 2006. Artifacts and pitfalls in diffusion MRI. J. Magn. Reson. Imag. 24, 478–488. https://doi.org/10.1002/jmri.20683.

Lebel, C., Walker, L., Leemans, A., Phillips, L., Beaulieu, C., 2008. Microstructural maturation of the human brain from childhood to adulthood. Neuroimage 40, 1044–1055. https://doi.org/10.1016/j.neuroimage.2007.12.053.

Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. In: Proc. 17th Sci. Meet. Int. Soc. Magn. Reson. Med., vol. 17, p. 3537.

Leemans, A., Jones, D.K., 2009. The *B* -matrix must be rotated when correcting for subject motion in DTI data. Magn. Reson. Med. 61, 1336–1349. https://doi.org/10.1002/mrm.21890.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2010. Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. 11, 19–60.

Mangin, J.-F., Poupon, C., Clark, C., Le Bihan, D., Bloch, I., 2002. Distortion correction and robust tensor estimation for MR diffusion imaging. Med. Image Anal. 6, 191–198. https://doi.org/10.1016/S1361-8415(02)00079-8.

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. Med. Image Comput. Comput. Assist. Interv. 9349, 12–19. https://doi.org/10.1007/978-3-319-24553-9_2.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C.E., Morey, R.A., Flashman, L.A., George, M.S., McAllister, T.W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R.D., Coleman, M.J., Kubicki, M., Westin, C.F., Stein, M.B., Shenton, M.E., Rathi, Y., 2016. Inter-site and inter-scanner diffusion MRI data harmonization. Neuroimage 135, 311–323. https://doi.org/10.1016/j.neuroimage.2016.04.041.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Karmacharya, S., Grant, G., Marx, C.E., Morey, R.A., Flashman, L.A., George, M.S., McAllister, T.W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R.D., Coleman, M.J., Kubicki, M., Westin, C.-F., Stein, M.B., Shenton, M.E., Rathi, Y., 2017. Multi-site harmonization of diffusion MRI data in a registration framework. Brain Imag. Behav. 1–12. https://doi.org/10.1007/s11682-016-9670-y.

Nunes, R.G., Hajnal, J.V., Golay, X., Larkman, D.J., 2006. Simultaneous slice excitation and reconstruction for single shot EPI. In: Proc Intl Soc Mag Reson Med, p. 293.

Perrone, D., Aelterman, J., Pižurica, A., Jeurissen, B., Philips, W., Leemans, A., 2015. The effect of Gibbs ringing artifacts on measures derived from diffusion MRI. Neuroimage 120, 441–455. https://doi.org/10.1016/j.neuroimage.2015.06.068.

Pierpaoli, C., 2010. Artifacts in diffusion MRI. In: Diffusion MRI. Oxford University Press, pp. 303–318. https://doi.org/10.1093/med/9780195369779.003.0018.

Pohl, K.M., Sullivan, E.V., Rohlfing, T., Chu, W., Kwon, D., Nichols, B.N., Zhang, Y., Brown, S.A., Tapert, S.F., Cummins, K., Thompson, W.K., Brumback, T., Colrain, I.M., Baker, F.C., Prouty, D., De Bellis, M.D., Voyvodic, J.T., Clark, D.B., Schirda, C., Nagel, B.J., Pfefferbaum, A., 2016. Harmonizing DTI measurements across scanners to examine the development of white matter microstructure in 803 adolescents of the NCANDA study. Neuroimage 130, 194–213. https://doi.org/10.1016/J.NEUROIMAGE.2016.01.061.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 22, 400–407.

Rudrapatna, S.U., Parker, G.D., Roberts, J., Jones, D.K., 2018. Can we correct for interactions between subject motion and gradient-nonlinearity in diffusion MRI? Proc. Int. Soc. Mag. Reson. Med. 1206.

Sairanen, V., Leemans, A., Tax, C.M.W., 2018. Fast and accurate Slicewise OutLIer Detection (SOLID) with informed model estimation for diffusion MRI data. Neuroimage 181, 331–346. https://doi.org/10.1016/J.NEUROIMAGE.2018.07.003.

Setsompop, K., Kimmlingen, R., Eberlein, E., Witzel, T., Cohen-Adad, J., McNab, J.A., Keil, B., Tisdall, M.D., Hoecht, P., Dietz, P., Cauley, S.F., Tountcheva, V., Matschl, V., Lenz, V.H., Heberlein, K., Potthast, A., Thein, H., Van Horn, J., Toga, A., Schmitt, F., Lehne, D., Rosen, B.R., Wedeen, V., Wald, L.L., 2013. Pushing the limits of in vivo diffusion MRI for the human connectome project. Neuroimage 80, 220–233. https://doi.org/10.1016/j.neuroimage.2013.05.078.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1874–1883. https://doi.org/10.1109/CVPR.2016.207.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155. https://doi.org/10.1002/hbm.10062.

St-Jean, S., Coupé, P., Descoteaux, M., 2016. Non Local Spatial and Angular Matching: enabling higher spatial resolution diffusion MRI datasets through adaptive denoising. Med. Image Anal. 32, 115–130. https://doi.org/10.1016/J.MEDIA.2016.02.010.

St-Jean, S., Viergever, M., Leemans, A., 2017. A unified framework for upsampling and denoising of diffusion MRI data. In: 25th Annual Meeting of ISMRM, p. 3533.

Tanno, R., Worrall, D.E., Ghosh, A., Kaden, E., Sotiropoulos, S.N., Criminisi, A., Alexander, D.C., 2017. Bayesian Image Quality Transfer with CNNs: Exploring Uncertainty in dMRI Super-resolution, pp. 611–619. https://doi.org/10.1007/978-3-319-66182-7_70.

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. Magn. Reson. Med. 73 https://doi.org/10.1002/mrm.25165.

Tax, C.M.W., Vos, S.B., Leemans, A., 2016. Checking and Correcting DTI Data, Diffusion Tensor Imaging: a Practical Handbook. https://doi.org/10.1007/978-1-4939-3118-7_7.

Thomas, C., Sadeghi, N., Nayak, A., Trefler, A., Sarlls, J., Baker, C.I., Pierpaoli, C., 2018. Impact of time-of-day on diffusivity measures of brain tissue derived from diffusion tensor imaging. Neuroimage 173, 25–34. https://doi.org/10.1016/J.NEUROIMAGE.2018.02.026.

Tibshirani, R.J., Taylor, J., 2012. Degrees of freedom in lasso problems. Ann. Stat. 40, 1198–1232. https://doi.org/10.1214/12-AOS1003.

Tournier, J.-D., Mori, S., Leemans, A., 2011. Diffusion tensor imaging and beyond. Magn. Reson. Med. 65, 1532–1556. https://doi.org/10.1002/mrm.22924.

Venkatraman, V.K., Gonzalez, C.E., Landman, B., Goh, J., Reiter, D.A., An, Y., Resnick, S.M., 2015. Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths. Neuroimage 119, 406–416. https://doi.org/10.1016/J.NEUROIMAGE.2015.06.078.

Veraart, J., Fieremans, E., Jelescu, I.O., Knoll, F., Novikov, D.S., 2016a. Gibbs ringing in diffusion MRI. Magn. Reson. Med. 76, 301–314. https://doi.org/10.1002/mrm.25866.

Veraart, J., Fieremans, E., Novikov, D.S., 2016b. Diffusion MRI noise mapping using random matrix theory. Magn. Reson. Med. 76, 1582–1593. https://doi.org/10.1002/mrm.26059.

Veraart, J., Novikov, D.S., Christiaens, D., Ades-aron, B., Sijbers, J., Fieremans, E., 2016c. Denoising of diffusion MRI using random matrix theory. Neuroimage 142, 394–406. https://doi.org/10.1016/J.NEUROIMAGE.2016.08.016.

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 3.0 T scanners. Neuroimage 51, 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., 2016. The importance of correcting for signal drift in diffusion MRI. Magn. Reson. Med. https://doi.org/10.1002/mrm.26124.

Zhu, T., Hu, R., Qiu, X., Taylor, M., Tso, Y., Yiannoutsos, C., Navia, B., Mori, S., Ekholm, S., Schifitto, G., Zhong, J., 2011. Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study. Neuroimage 56, 1398–1411. https://doi.org/10.1016/J.NEUROIMAGE.2011.02.010.