Contents lists available at ScienceDirect

NeuroImage

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

Incorporating outlier information into diffusion-weighted MRI modeling for robust microstructural imaging and structural brain connectivity analyses

Viljami Sairanen[a,b,c,*], Mario Ocampo-Pineda[b], Cristina Granziera[c], Simona Schiavi[b], Alessandro Daducci[b]

[a] BABA Center, Pediatric Research Center, Department of Clinical Neurophysiology, Children's Hospital, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
[b] Department of Computer Science, University of Verona, Verona, Italy
[c] Translational Imaging in Neurology, Department of Medicine and Biomedical Engineering, University Hospital Basel and University of Basel, Neurologic Clinic and Policlinic, Basel, Switzerland

ARTICLE INFO

Keywords:
Magnetic resonance imaging
Diffusion
Tractography
Robust
Filtering
Commit
Outlier
Imputation
Structural connectivity
Microstructure
Movement
Weighted modeling

ABSTRACT

The white matter structures of the human brain can be represented using diffusion-weighted MRI tractography. Unfortunately, tractography is prone to find false-positive streamlines causing a severe decline in its specificity and limiting its feasibility in accurate structural brain connectivity analyses. Filtering algorithms have been proposed to reduce the number of invalid streamlines but the currently available filtering algorithms are not suitable to process data that contains motion artefacts which are typical in clinical research. We augmented the Convex Optimization Modelling for Microstructure Informed Tractography (COMMIT) algorithm to adjust for these signals drop-out motion artefacts. We demonstrate with comprehensive Monte-Carlo whole brain simulations and in vivo infant data that our robust algorithm is capable of properly filtering tractography reconstructions despite these artefacts. We evaluated the results using parametric and non-parametric statistics and our results demonstrate that if not accounted for, motion artefacts can have severe adverse effects in human brain structural connectivity analyses as well as in microstructural property mappings. In conclusion, the usage of robust filtering methods to mitigate motion related errors in tractogram filtering is highly beneficial, especially in clinical studies with uncooperative patient groups such as infants. With our presented robust augmentation and open-source implementation, robust tractogram filtering is readily available.

1. Introduction

Diffusion-weighted magnetic resonance imaging (dMRI) of the human brain (Basser et al., 1994) has various applications ranging from early clinical stroke diagnostics (Horsfield and Jones 2002) to investigations of the microstructural properties of the tissue (Alexander et al., 2019; Novikov et al., 2019) and structural brain connectivity mapping (Griffa et al., 2013; Delettre et al., 2019; Zhang et al., 2021) The latter two are gaining popularity in clinical research (Kamiya et al., 2020) to investigate various brain diseases and neurological conditions of adults (Fieremans et al., 2013; Benitez et al., 2014) and development of the growing brain in children and adolescents (Genc et al., 2017; Huber et al., 2019). Furthermore, with the latest advances in automatic brain segmentation with tools like Infant FreeSurfer (Zöllei et al., 2020), it is likely that the amount of brain connectivity studies of infants (Kunz et al., 2014; Pannek et al., 2018; Pecheva et al., 2019) will grow in the near future too.

The clinical dMRI research comes with its own challenges to solve, with one most difficult being the patient motion. The subject motion can be unavoidable when imaging infants or patients in discomfort or pain, resulting in complex missing data problems (Sairanen et al., 2017, 2018). In short, rapid subject motion can result in slicewise signal dropout artefacts. For readers interested in why this happens, we recommend the section "Origin of the dropout" by Andersson et al. (2016). Therefore, the processing of the motion-corrupted images requires specialized algorithms and robust methods to minimize motion induced bias in the results. While robust modeling has been considered in the contexts of diffusion and kurtosis tensor estimations (Chang et al., 2005; Chang et al., 2012; Tax et al., 2015) as well as in higher order models (Pannek et al., 2012) that could be used for tractography purposes, it has

[*] Corresponding author at: BABA Center, Pediatric Research Center, Department of Clinical Neurophysiology, Children's Hospital, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
E-mail addresses: viljami.sairanen@helsinki.fi, viljami.sairanen@hus.fi (V. Sairanen).

https://doi.org/10.1016/j.neuroimage.2021.118802.
Received 10 June 2021; Received in revised form 1 November 2021; Accepted 9 December 2021
Available online 10 December 2021.

not been investigated thoroughly in the context of the brain structural connectivity analyses.

Structural brain connectivity analyses are based on the rapidly developing dMRI tractography (Basser et al., 2000) algorithms that represent the brain white matter structures with streamlines. These streamlines can be used to investigate which gray matter regions might have a structural link. In general, the tractography algorithms are sensitive but they lack specificity and they find great number of false streamlines connections (Thomas et al., 2014; Maier-Hein et al., 2017). This means that two gray matter regions could be linked by tractography streamlines despite that the brain tissue does not form a true structural link. This is a known issue in structural connectivity analyses (Drakesmith et al., 2015; Zalesky et al., 2016; Yeh et al., 2020) to which tractogram filtering has been proposed as one solution. Tractogam filtering can be achieved with different approaches (Zhang et al., 2021), one being the Convex Optimization Modelling for Microstructure Informed Tractography (COMMIT) (Daducci et al., 2015) which we will use in this study to demonstrate possible effects of subject motion to the filtering and microstructural mapping as well as how it can be accounted and corrected for.

There are three alternative post-scan approaches to address outliers that are caused by the subject motion. The first approach is to find outliers in dMRI data manually or automatically with statistical methods or deep-learning and simply exclude the artefactual dMRI data or even the whole subject from the analysis (Oguz et al., 2014; Samani et al., 2019). The second approach is to use a model to predict what the measurements should look like, locate the outliers based on differences to model predictions and replace them with these predictions if differences are deemed large enough (Lauzon et al., 2013; Andersson et al., 2016). The third approach is to detect the outliers, but instead of replacing or completely excluding them, their weight is reduced in all subsequent model estimation steps (Sairanen et al., 2018).

Manual outlier detection can be laborious and excluding whole subjects from clinical studies with relatively small number of participants might not be the optimal choice. The outlier replacement approach relies on the quality and robustness of the chosen model and method to represent the measured dMRI signal. If multiple dMRI measurements are corrupted by motion artefacts, this initial modeling and prediction step can fail altogether (Sairanen et al., 2018). Even in the best case, the replaced data points are simply interpolations based on the chosen model and the data points used in the modeling therefore it cannot increase the available information but leads to increased error propagation due to subsequent model fittings. The third approach, on the contrary, enables quantifying the amount of the motion corrupted data and versatile subsequent modeling and analysis options therefore being optimal for our purposes. In Discussion section “Robust modeling or outlier replacement”, we provide further reasoning why we promote the use of robust methods over replacement in dMRI.

While weighted and robust modeling has been implemented before, they have mostly been used outside the scope of tractogram filtering. For example, in diffusion tensor modeling weighted linear least squares is typically the fastest and most robust estimator (Veraart et al., 2013; Tax et al., 2015; Sairanen et al., 2018). Robust modeling has been proposed for higher order models as well (Pannek et al., 2012). In the context of tractogram filtering, weighted cost functions have been introduced earlier in e.g., SIFT (Smith et al., 2013, 2015), but it has only been evaluated with voxels affected by partial voluming. SIFT algorithm states that their ’processing mask’ is ’the square of the estimated white matter partial volume fraction’ - which indeed should be beneficial in the case of partial voluming. However, the approach in SIFT does not account for outliers that are randomly occurring in the measurements as our newly proposed augmentation to COMMIT does.

In this work, we propose a robust augmentation to the COMMIT algorithm (Daducci et al., 2015) that accounts for the unreliability of the original measurements. We detail the theoretical changes to the algorithm as well as provide open-source code[1] of its implementation. We refer to this new method as COMMIT_outlier throughout this manuscript. To evaluate the method, we use the data from the Human Connectome Project (HCP) (Van Essen et al. 2013) as a base for thorough Monte-Carlo simulations which emulate various motion induced artefacts in synthetic but realistic whole brain data. Synthetic data provides the necessary baseline that can be used to isolate the bias arising from subject motion from noise effects in structural connectivity analyses as well as how well motion artefacts can be amended using our robust augmentation. In the context of this study, the measurement unreliability is associated with outliers due to subject motion. However, it can readily be utilized to correct for measurements that are affected by partial voluming, as our preliminary results have demonstrated earlier (Sairanen et al., 2021).

2. Material and methods

2.1. Implementation

We augmented the original cost function of COMMIT (Daducci et al., 2015) with a voxelwise weighting factor **W** that we used to down weight measurements that have decreased reliability due to subject motion or any other reason. The original COMMIT is based on a minimization of the difference between the original measurements and a forward model prediction. The forward model prediction is calculated by fitting a chosen microstructural model for each streamline in every voxel. COMMIT assigns a weight to each streamline that tells how much that streamline contributes to the predicted signal. These streamline contribution weights are iteratively updated until the difference between the measurements and this prediction converges to a minimum. Any streamline with contribution of zero is then removed as an implausible streamline (i.e., not compatible with the measured signal).

If part of the measurements are artefactual due to subject motion or any other reason, the original COMMIT algorithm could converge to an incorrect solution. To avoid this and decrease the impact of these artefactual measurements, we propose the robust cost function shown in Eq. (1). The weighting factor **W** is used to multiply the difference between the original measurements **y** and the product of model design matrix **A** and estimated model coefficients **x** in the minimization problem. Our proposed idea is further illustrated in Fig. 1 with a simple toy example. In future, these reliability weights **W** could be iteratively updated along with the model coefficients **x** to help in estimating model coefficients in voxels that do not fit to the chosen model perfectly due to heart beat related pulsation or other uncertainties.

$$\underset{\hat{x}\geq 0}{argmin} \|W(A\hat{x}-y)\|_2^2. \tag{1}$$

The robustly weighted cost function in Eq. (1) is intended to be used with outlier detection with tools such as SOLID (Sairanen et al., 2018). SOLID detects slicewise outliers based on robust statistical analysis of the original dMRI data and can be used either to exclude outliers or down weight them depending on how strong outliers are. This down weighting scheme is likely a better option to outlier replacement that is proposed in earlier studies (Lauzon et al., 2013; Andersson et al., 2016). If the outlier is replaced with a prediction from a tensor or a gaussian model, then COMMIT would try to minimize the difference from those model predictions to its own model prediction. Since these models can be different and therefore capture different details of the dMRI signal, it is more straightforward to use robust modeling with the proposed weighted cost function. For interested readers, we provide more reasoning for this claim in the Discussion section “Robust modeling or outlier replacement”.

[1] https://github.com/daducci/COMMIT.

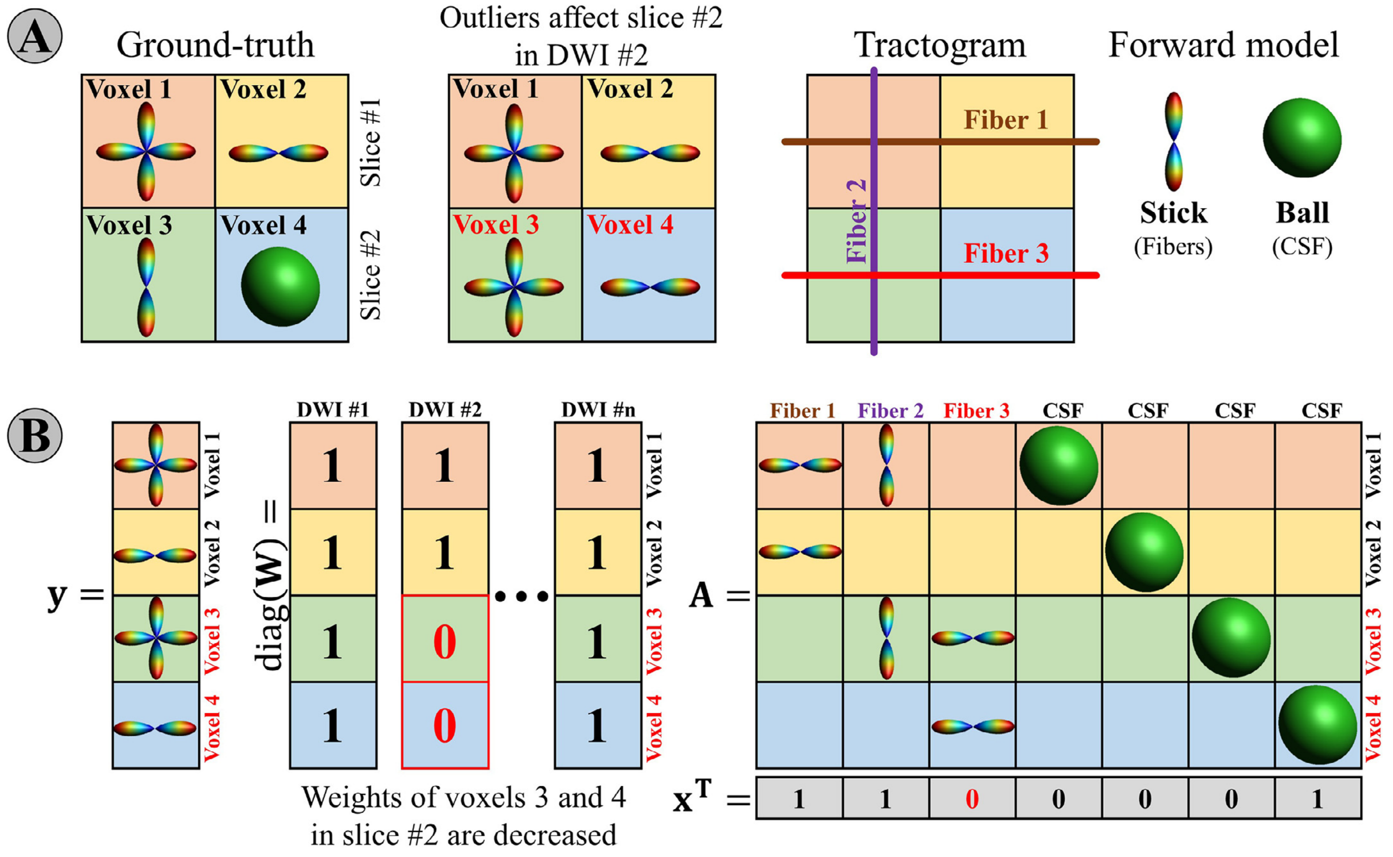

Fig. 1. A toy example to illustrate the augmentation of COMMIT with the measurement reliability-based weighing. (**A**) A synthetic phantom consists of two slices both consisting of two voxels. For visualization purposes the 4th dimension i.e., diffusion weighted signals are omitted and observed signal is visualized using a fiber orientation distribution (FOD). To illustrate the subject motion artefact, the second slice in DWI #2 is affected by slicewise artefact and FOD of corresponding voxels 3 and 4 is biased. In tractography, this is seen as an implausible streamline Fiber 3. (**B**) Data is vectorized to visualize the linear model problem. Vector **y** contains the simulated signals, diagonal of matrix **W** contains the reliability weights that are decreased for voxels 3 and 4 in DWI #2, **A** is the design matrix that depicts the modeled compartments (e.g., stick for streamlines and ball for isotropic compartments), and **x** contains the streamline wise contributions. With the successful downweighing, the contribution of Fiber 3 is set to zero and the implausible streamline is thus removed.

2.2. Simulations

To investigate the outlier effect on the tractogram filtering, we developed a comprehensive Monte-Carlo simulation pipeline delineated in Fig.2. Simulations were based on T1-weighted and dMRI data from the HCP subject 103,818 which were processed with current state-of-the-art methods (Van Essen et al. 2013). We do not expect or imply that this ground truth connectivity matrix depicted in Fig. 3 would represent the true structural connections in a human brain. It simply provides us the necessary ground truth connectivity that we can use to evaluate the noise and outlier effects in the Monte-Carlo simulations with more realistic picture of the whole brain than typical fiber phantoms as it contains realistic brain structures such as kissing or crossing fibers as well as the modeled partial voluming effects.

2.2.1. Ground truth data

We segmented the T1-weighted HCP data with FreeSurfer (Fischl 2012) to obtain 85 regions-of-interests (ROIs) based on the Desikan et al. (2006) atlas. Instead of the full brainstem, we used only its inferior part of medulla as the last ROI. We used these brain segments to compute the ground truth connectivity matrix as well as to ensure that we used only the connecting streamlines in our analyses.

To calculate a whole brain tractogram from the HCP dMRI data, we used the anatomically constrained probabilistic tractography (iFOD2) (Tournier et al., 2010; Smith et al., 2012) implemented in MRTrix3 software (Tournier et al., 2019). We used the white matter mask as a seed region for three million streamlines. The tracking parameters were: step size 0.5, turning angle 45°, min length value 5, max length 250, cutoff value 0.05, trials number 1000. Finally, we removed all non-connecting streamlines based on the 85 ROI segmentation of T1-image.

For ground truth tractogram filtering, we used the original COMMIT (Daducci et al., 2015) because data did not contain slicewise outliers. We chose the stick-zeppelin-ball (SZB) as the forward model with following parameters: $1.7 \cdot 10^{-3} mm^2/s$ for parallel stick and zeppelin diffusivities, $0.61 \cdot 10^{-3} mm^2/s$ for perpendicular zeppelin diffusivity, and two ball compartments of $1.7 \cdot 10^{-3} mm^2/s$ and $3.0 \cdot 10^{-3} mm^2/s$to account for partial volumeing with gray matter as well as in cerebrospinal fluid (Panagiotaki et al., 2012).

The filtered tractogram was used to form the ground truth connectivity matrix with the information from T1-segmentation (Fig. 3). The network edges in the ground truth connectivity matrix were defined as the sum of the COMMIT streamline weights multiplied by the length of the tract and normalized by the average tract length between each gray matter parcellation as was done in (Schiavi et al., 2020). We combined this information with the final streamline contributions to form the synthetic whole brain prediction of dMRI data using the HCP's three-shell gradient scheme. This produced 270 noise free diffusion-weighted whole brain images that we used as a ground truth for our Monte-Carlo simulations.

2.2.2. Monte-Carlo data

Our Monte-Carlo simulations were based on the ground truth synthetic whole brain dMRI data obtained from HCP subject. We split the simulations into two groups: Baseline and Test. Baseline group provides

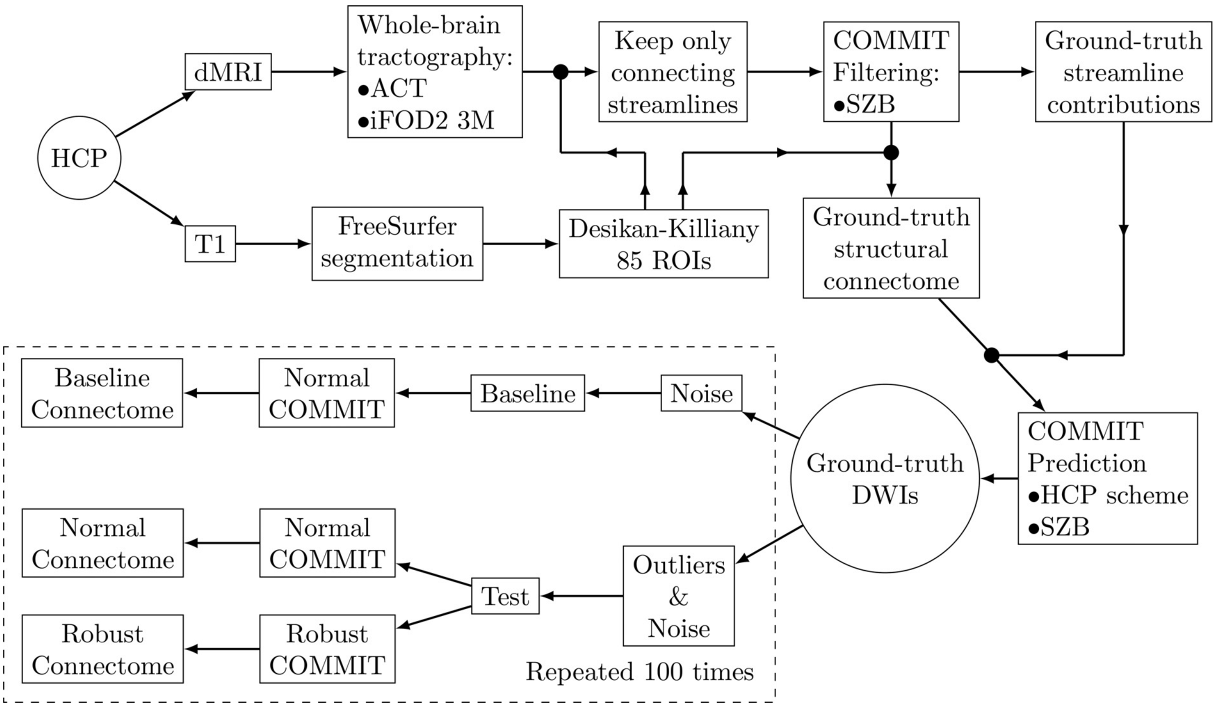

Fig. 2. A flowchart describing how the whole-brain simulations were obtained from the HCP dataset. The dMRI and T1-weighted data were used to obtain the ground truth connectome from which the ground truth dMRI signals were predicted using normal COMMIT forward modeling. The ground truth data was used to perform 100 Monte-Carlo simulations to evaluate the effects of noise and outliers to the structural brain connectome. The Monte-Carlo iteration setups shown inside the dashed rectangle were repeated for outlier percentages 5% and 10% both with the uniform and clustered outlier schemes.

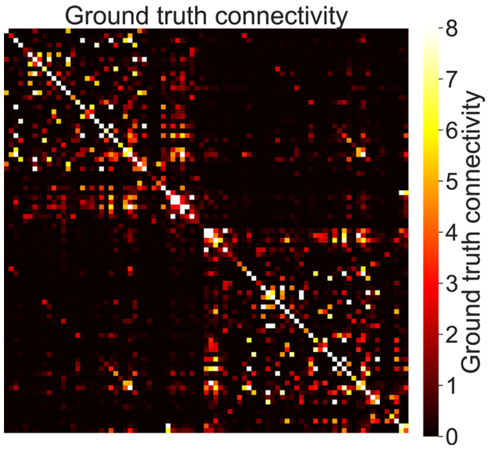

Fig. 3. The ground truth connectivity matrix used in this study was based on one subject. While, this connectivity matrix might not represent the real human brain connections, it provided the necessary ground truth control for our Monte-Carlo simulations. Deviations from this network in the simulations would be due to noise, outliers, or both.

the means to evaluate the pure noise effects on the connectome whereas Test group provides the means to isolate and evaluate the outlier effects. Network analyzes used for Test group were identical with the ground truth case.

In Baseline group, random Rician noise was added before repeating the normal COMMIT filtering with the original non-filtered but connecting streamlines. The Rician noise had signal-to-noise ratio of 20 based on the non-diffusion weighted signal which is roughly similar with signal-to-noise ratios in clinical research. We used the same filtering parameters that were used to form the ground truth data. This process was repeated to obtain 100 whole brain baseline images and connectomes.

In Test group, outliers were introduced to the data before adding the same Rician noise that was used for the Baseline group. Test group was filtered with both the normal COMMIT as well as the proposed robust COMMIT_outlier using the same streamlines and parameters that were used for the Baseline group. This process was repeated to obtain 100 whole brain test images with outliers and corresponding connectomes from normal and robust filtering methods.

The outlier selection for the Test group was done with two different schemes by replacing axial slices with signal decrease outliers in an interleaved manner to 5% and 10% of the dMRI data per shell. The first scheme represented the worst possible situation where outliers were clustered in the q-space (e.g., Fig. 4) whereas the second scheme represented the best possible situation where outliers were uniformly placed in the q-space based on their electrostatic repulsion (Sairanen et al., 2017). Futher details why we did not use purely random selection of outliers is provided in Discussion section “Robust modeling or outlier replacement”.

2.2.3. Statistical analysis

We investigated global brain connectivity as well as individual network edges using analysis of variance (ANOVA) accompanied by Tukey’s honestly significant difference (HSD) test and non-parametric Friedman’s test accompanied by two-sample Kolmogorov-Smirnov tests. The reason for having these different test statistics is that outliers can lead to skewed and long tailed distributions that might not be correctly investigated solely by parametric tests. It should be noted that the added Rician noise has a non-zero positive mean value. This means that all groups are likely shifted to some direction from the ground truth prediction. This is the reason, why the Baseline group is needed as that is affected only by noise and can be used to isolate corresponding shifting effect.

While we report p-values from these tests, we argue that the effect sizes are more interesting as they describe how different the tested groups are. The effect sizes are measured using Cohen’s D for paramet-

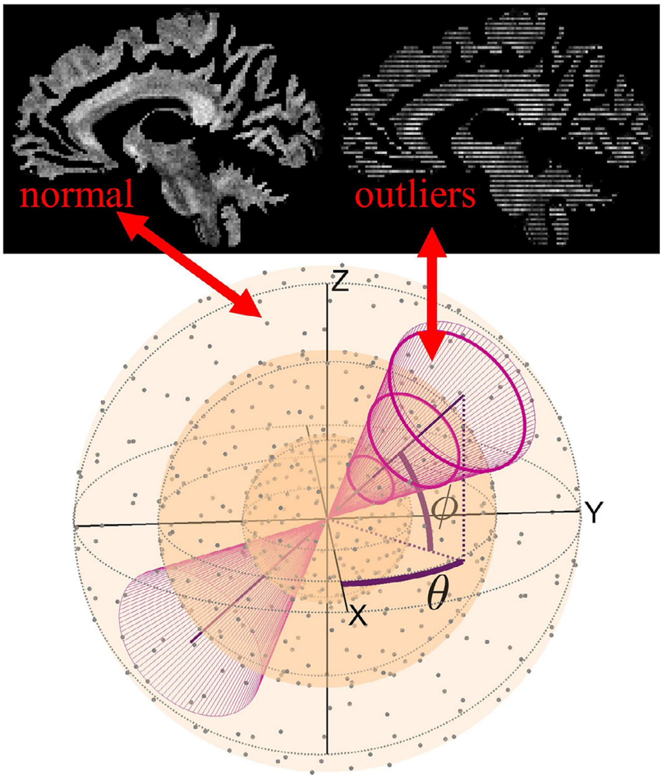

Fig. 4. An illustration how clustered outliers were selected from all the gradient directions. The three shells used in the HCP gradient scheme are shown with the transparent spheres and gradient directions with black dots. The initial direction (θ,φ) of was selected randomly after which the opening angle of the cone was increased until wanted number of outliers from each shell remained inside it. This approach ensured the maximal gap in the q-space sampling and the chance to find error prone schemes.

ric tests and Kolmogorov-Smirnov statistic for non-parametric tests. The test statistics we employ are widely used and they provide information about average differences and differences in the shapes of the Monte-Carlo simulated distributions. For details about these tests, we recommend any textbook that covers parametric and non-parametric statistics such as Sheskin (2004) handbook. Having these two different tests seemed necessary during designing this study. Of course, as we did multiple tests, we also performed multiple comparison correction to all the p-values in pair-wise tests. In total, there were 80 different tests and we corrected for them using Benjamini-Hochberg False Discovery Rate (FDR) test (Benjamini and Hochberg, 1995) with alpha 0.05. However, as seen from the results later, neither mean based or distribution shape-based analysis might not be sufficient to provide absolute conclusions.

2.3. In vivo measurements

2.3.1. Infant data

We obtained preliminary data from an on-going infant study to evaluate our method with in vivo measurements. T1-weighted image and dMRI data were obtained with 3T MRI Siemens Skyra system (Erlangen, Germany) with a 32-channel head coil. The dMRI acquisition consisted of 13 non-diffusion weighted images that were interspersed between 60 diffusion-weighted images with b-value of $750 s/mm^2$ and 74 diffusion-weighted images with b-value of $1800 s/mm^2$each with uniquely oriented gradients. Bipolar gradient scheme was used to minimize geometrical distortions due to eddy currents. The image resolution was isotropic 2 mm with 80×80×44 imaging matrix. The in-plane acceleration factor was 2 (SENSE) and multi-band acceleration factor was 2. Only anterior-posterior phase encoded images were acquired as the reverse phase encoding required manual adjustment during the scan which was deemed infeasible at the corresponding clinical scan environment. The use of infant data in this work was approved by the relevant Ethics Committee of the Helsinki University Hospital.

2.3.2. Infant analyses

We used ExploreDTI (Leemans et al., 2009) with SOLID-plugin (Sairanen et al., 2018) to simultaneously detect slicewise outliers and to correct for subject motion and eddy currents as well as registered the data to anatomical T1-image to correct for geometrical distortions. Additionally, we used Gibbs ringing correction (Perrone et al., 2015). We did not correct for signal drift (Vos et al. 2016) as it was not observed in the measurements.

Processing of this data was limited to specific computers in the hospital network which prevented memory demanding tasks such as segmentation with Infant Freesurfer (Zöllei et al., 2020). Problematically, the T1-image contrast of this subject was not suitable for white and gray matter segmentations using traditional options. This, unfortunately, prevented us from performing full network analyses on the infant dataset as there was no reliable way to perform gray matter segmentation and we had to content to a simpler analysis that consisted of comparing signal fraction maps between normal and robust filtering method. We obtained a WM mask from multi-shell multi-tissue constrained spherical deconvolution (Jeurissen et al., 2014) implemented in MRTrix3 (Tournier et al., 2019) and used that as a seed mask for probabilistic whole-brain tractography (iFOD2) (Tournier et al., 2010) to generate three million streamlines.

We filtered the generated streamlines with normal COMMIT (Daducci et al., 2015) and the proposed robust COMMIT_outlier to evaluate the improvements in the overall fit from root mean squared error (RMSE) maps as well as to see the impact of outliers in intracellular and isotropic signal fractions. We used the stick-ball model for both filtering methods with the following parameters: $1.7 \cdot 10^{-3} mm^2/s$ for parallel signal diffusivity, and $1.7 \cdot 10^{-3} mm^2/s$and $3.0 \cdot 10^{-3} mm^2/s$ for the isotropic signal diffusivities.

3. Results

3.1. Simulations

We investigated the effects of noise to the structural brain connectivity by comparing Baseline group. Test groups (COMMIT and proposed robustly weighted COMMIT_outlier) could not be directly compared to the ground truth due to Rician noise bias. With the Rician noise bias we imply the effect that adding noise with non-zero mean (Gudbjartsson and Patz 1995) to data leads to a shift in overall baseline. Therefore, outlier effects were investigated by comparing Test groups to Baseline group. We evaluated these differences in both global connectivity matrix score as well as in network edge individually.

3.1.1. Global connectivity

The global connectivity difference was defined as an average absolute difference between the element's upper triangle of the connectivity matrices from Monte-Carlo groups and the corresponding ground truth values. The results of this comparison calculated are shown in Fig. 5 with all violin plots being based on 100 Monte-Carlo simulations each. The noise effect on the global connectivity (Baseline) is shown with the first violin from the left, the uniform outlier effect is shown in the middle, and the clustered outlier effect is shown on the right. The percentage of outliers (5% or 10%) is shown on different sides of each violin.

Both, Baseline and robust COMMIT_outlier produced similar global results with differences ranging from $3.5 \cdot 10^{-4}$to $4.0 \cdot 10^{-4}$. This demonstrates that on average, the proposed robust filtering method is capable to mitigate the outlier effect. On the contrary, the results from normal COMMIT ranged from $3.25 \cdot 10^{-4}$ to $6.5 \cdot 10^{-4}$demonstrating that outliers

Table 1
Summary of the parametric and non-parametric test results. Bolded p-values indicate statistically significant findings with FDR based correction for multiple comparisons using 0.05 alpha.

		Global, 5% outliers				Global, 10% outliers				CST, 5% outliers				CST, 10% outliers			
		ANOVA		Friedman		ANOVA		Friedman		ANOVA		Friedman		ANOVA		Friedman	
		score	p	score	p	score	p	score	p	score	p	score	p	score	p	score	p
		124.51	< 0.05	224.10	< 0.05	281.80	< 0.05	290.88	< 0.05	12.93	< 0.05	152.86	< 0.05	1.81	0.13	115.26	< 0.05
Groups		D	HSD	K-S stat.	K-S	D	HSD	K-S stat.	K-S	D	HSD	K-S stat.	K-S	D	HSD	K-S stat.	K-S
Baseline	Uniform COMMIT	2.06	**0.00**	0.73	**0.00**	3.82	**0.00**	0.96	**0.00**	2.71	**0.00**	0.89	**0.00**	2.11	0.13	0.75	**0.00**
Baseline	Uniform COMMIT_r	0.22	0.72	0.15	0.21	0.46	0.43	0.24	**0.01**	0.29	0.90	0.14	0.28	0.28	0.90	0.14	0.28
Baseline	Cluster COMMIT	1.76	**0.00**	0.69	**0.00**	1.89	**0.00**	0.78	**0.00**	0.29	**0.01**	0.48	**0.00**	0.08	0.88	0.52	**0.00**
Baseline	Cluster COMMIT_r	0.18	0.83	0.13	0.37	0.31	0.75	0.23	**0.01**	0.22	0.90	0.15	0.21	0.41	0.90	0.22	**0.02**
Uniform COMMIT	Uniform COMMIT_r	2.28	**0.00**	0.75	**0.00**	3.28	**0.00**	0.93	**0.00**	2.53	**0.00**	0.84	**0.00**	1.94	0.21	0.72	**0.00**
Uniform COMMIT	Cluster COMMIT	0.61	**0.00**	0.44	**0.00**	3.23	**0.00**	0.98	**0.00**	0.25	0.05	0.46	**0.00**	0.13	0.59	0.52	**0.00**
Uniform COMMIT	Cluster COMMIT_r	2.18	**0.00**	0.77	**0.00**	3.42	**0.00**	0.93	**0.00**	2.49	**0.00**	0.84	**0.00**	1.68	0.27	0.62	**0.00**
Uniform COMMIT_r	Cluster COMMIT	1.89	**0.00**	0.71	**0.00**	2.05	**0.00**	0.82	**0.00**	0.25	0.05	0.48	**0.00**	0.06	0.90	0.52	**0.00**
Uniform COMMIT_r	Cluster COMMIT_r	0.03	0.90	0.13	0.37	0.15	0.90	0.10	0.70	0.05	0.90	0.11	0.58	0.16	0.90	0.18	0.08
Cluster COMMIT	Cluster COMMIT_r	1.85	**0.00**	0.72	**0.00**	2.00	**0.00**	0.81	**0.00**	0.26	0.04	0.48	**0.00**	0.05	0.90	0.52	**0.00**

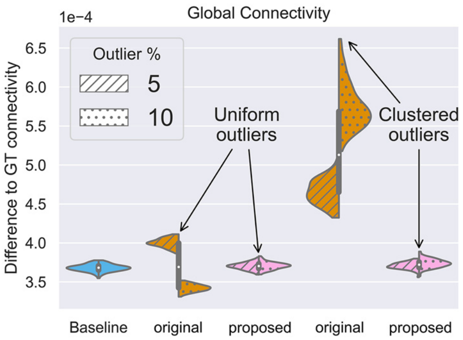

Fig. 5. Impact of noise (Baseline) and outliers (original and proposed) to the global structural brain connectivity. Left and right sides of the violins represent simulations with 5% and 10% outliers, respectively. The y-axis indicates the distance to the ground truth as an average absolute difference. The augmented COMMIT_outlier shown with the pink violins produced similar distributions with the Baseline in all cases whereas the original COMMIT shown with orange violins differs from the Baseline already in the 5% cases. As expected, the clustered outlier scheme produced the largest deviations with the highest variability in the original COMMIT distributions. Interestingly, the uniform outlier scheme resulted two different distributions for original COMMIT compared to the Baseline. This highlights the need for the robust processing as the exact effect of outliers can be very challenging to predict.

can have a much stronger effect than noise on the global connectivity values.

Parametric statistical analysis

The global connectivity differences with ANOVA detail that the group averages were statistically different with p-value less than 0.05. Tukey's HSD test results are shown in Table 1 along with all other statistical tests results. Statistically significant results after multiple comparison correction with FDR alpha 0.05 are shown with bolded p-values. The results depict that normal COMMIT had significantly different mean to both Baseline and COMMIT_outlier results and the effect sizes evaluated with Cohen's D were systematically larger. Importantly, differences between Baseline and COMMIT_outlier were not statistically significant with relatively small effect sizes. These effect sizes indicate that in our realistic simulations with 5% and 10% of outliers, the average bias caused by outliers quickly increases and compromises the connectivity analyses if data is not processed robustly.

Non-parametric statistical analysis

The Friedman's test reported also p-value less than 0.05 therefore providing additional support for the graphical analysis and ANOVA results. We applied the two-sample Kolmogorov-Smirnov test to detect which of the distributions were different. All these test results are reported in Table 1. Comparisons between Baseline and normal COMMIT were all statistically significant with large effect sizes whereas comparisons between Baseline and COMMIT_outlier were not statistically significant with 5% outliers. With 10% outliers non-parametric differences between Baseline and COMMIT_outlier were significant but the effect size remained small.

3.1.2. Network edges

We investigated the network edge-wise differences between the Monte-Carlo connectivity matrices with parametric and non-parametric statistics as complementary information to the global results. The three violin plots in Fig. 6 depict the connectivity values from medulla to the right precentral gyrus. These streamlines are visualised in Fig. 7 and are likely a part of the corticospinal tract and therefore a known true connection. The results of the parametric and non-parametric tests performed to this network edge are depicted in Table 1.

The noise effect results in a systematic over estimation of the connectivity strength as depicted by Baseline in Fig. 6. However, outliers have a more random effect depending on the affected dMRI measurements. This can either decrease or increase the connectivity strength and can counteract the noise effect. Therefore, group comparisons against Baseline were more meaningful than comparisons against the known ground truth value would be. For example, in this case the normal COMMIT produces an average connectivity strength that is closer to the ground truth than Baseline despite the distribution is wider.

Parametric statistical analyses

The connectivity-wise differences between Baseline and normal COMMIT as well as Baseline and robust COMMIT_outlier are shown in Fig. 8. The color map indicates the effect size measured with Cohen's D. Only elements that were deemed significantly different (p-value less

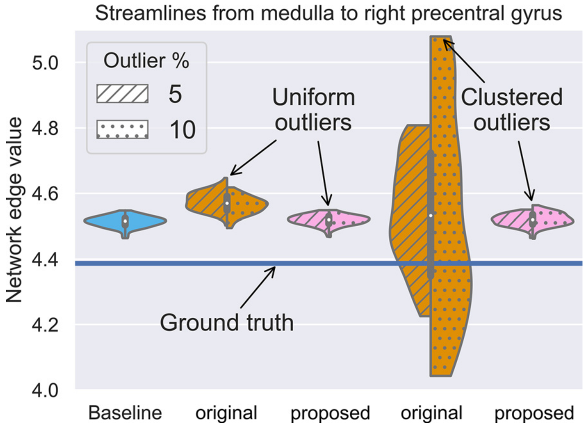

Fig. 6. Impact of noise (Baseline) and outliers (original and proposed) to one specific network edge that represents connections between medulla and right precentral gyrus. The y-axis indicates the strength of this edge. Left and right sides of the violins represent simulations with 5% and 10% outliers, respectively. The augmented COMMIT_outlier shown with the pink violins produced similar distributions with the Baseline in all cases whereas the original COMMIT shown with orange violins was heavily affected already in the 5% cases. As expected, the clustered outlier scheme produced the largest deviations with the highest variability in the original COMMIT distributions. The original COMMIT simulations with the clustered outlier scheme demonstrate why it is necessary to compare results against the Baseline instead of the noiseless ground truth as the outlier effect can surpass the noise effect. This can lead to the shown situation where the difference to ground truth on average would be smaller due to a very wide distribution.

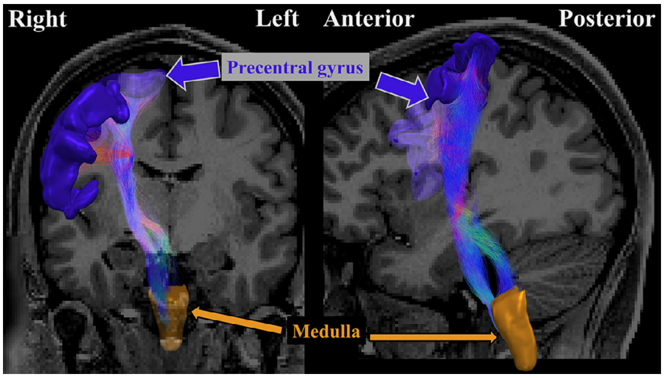

Fig. 7. Illustration of the network edge that connects medulla to right precentral gyrus. This edge or connection was selected for closer inspection as it forms a part of the corticospinal tract which is well known true connection in a healthy human brain. Shape of the tractogram is nearly vertical therefore perpendicular to the introduced slice-wise outliers in the axial plane.

than 0.05) based on ANOVA and Tukey's HSD were drawn. The comparison between Baseline and normal COMMIT resulted in more elements with significant differences than the comparison between Baseline and COMMIT_outlier. The effect sizes between Baseline and normal COMMIT ranged from 0 up to 3 indicating that outliers can have strong adverse effects on specific connectivity matrix elements. The overall smaller effect sizes between Baseline and robust COMMIT_outlier highlight that our augmentation is well capable to mitigate the outlier effects even on individual network edge level.

Non-parametric statistical analysis

The connectivity-wise distributional differences between Baseline and normal COMMIT as well as Baseline and robust COMMIT_outlier are shown in Fig. 9. The color map indicates the effect size measured with Kolmogorov-Smirnov statistic. Only elements that were deemed statistically significantly different (p-value less than 0.05) based on

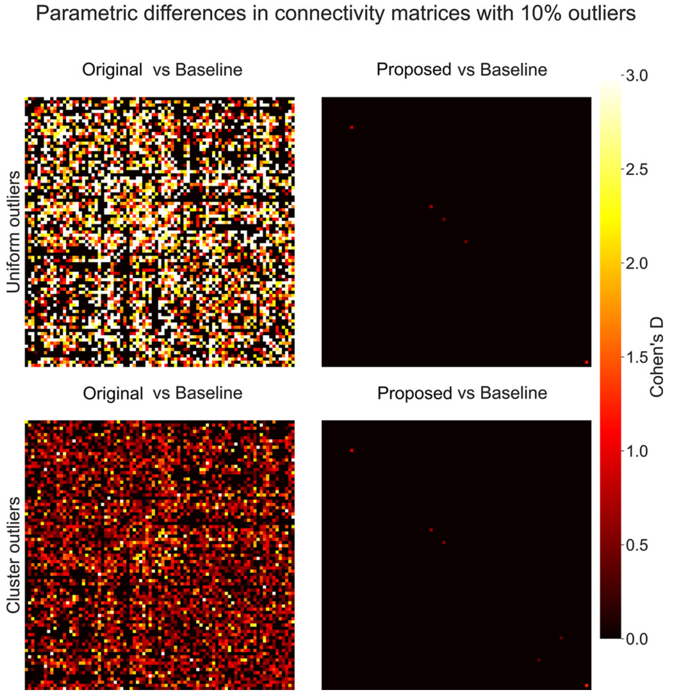

Fig. 8. Differences in the network edges due to outliers measured using parametric statistics. The color scale indicates the average effect size calculated using Cohen's D with unequal variances. The left and right columns show the difference from Baseline to original and proposed robust COMMIT. The top and bottom rows show the results from uniform and clustered outlier schemes. Robust augmentation clearly improves the COMMIT filtering if the data contains outliers as the effect sizes remain very small in all edges in both outlier schemes. While the uniform outlier scheme produced larger effect sizes for original COMMIT than the clustered, it can be easily explained because Cohen's D is inversely proportional to the sample variance which is very high in the clustered outlier schemes.

Kolmogorov-Smirnov tests were drawn. Similar to the parametric counterpart, the differences between Baseline and normal COMMIT were again more frequent than differences between Baseline and robust COMMIT_outlier. Also, the effect sizes between Baseline and normal COMMIT ranged from 0 to nearly 1 which is the maximum of the used statistic. This indicates that outliers can lead to very large distributional differences. The differences between Baseline and robust COMMIT_outlier remained relatively small with effect sizes ranging from 0 to 0.2.

3.2. In vivo measurements

Besides tractogram filtering, we calculated the intracellular and isotropic signal fractions using the COMMIT (Daducci et al., 2015) and the proposed robust COMMIT_outlier. Fig. 10 shows the results for outlier detection, RMSE, and signal fraction maps obtain from the infant data. On average, the amount of missing data i.e., how much confidence in fitting was decreased per slice position ranged from 5% to 19%.

The RMSE map of normal COMMIT was clearly affected by the outliers resulting in visible stripes in the image. On the contrary, the COMMIT_outlier RMSE map that describes the robust cost function does not have such stripes therefore the fitting is not affected by outliers. The difference RMSE map visualises the stripy pattern more prominently and ranges from 0 to 30%. The outlier effect on intracellular and isotropic signal fractions was less prominent in visual analysis i.e. less or no stripes. However, the difference between normal COMMIT and robust COMMIT_outlier depicts that the differences ranged from −10% to +10% even in regions that were less affected by outliers for intracellular signal fraction. For isotropic signal fraction the differences ranged from −7% to +7%.

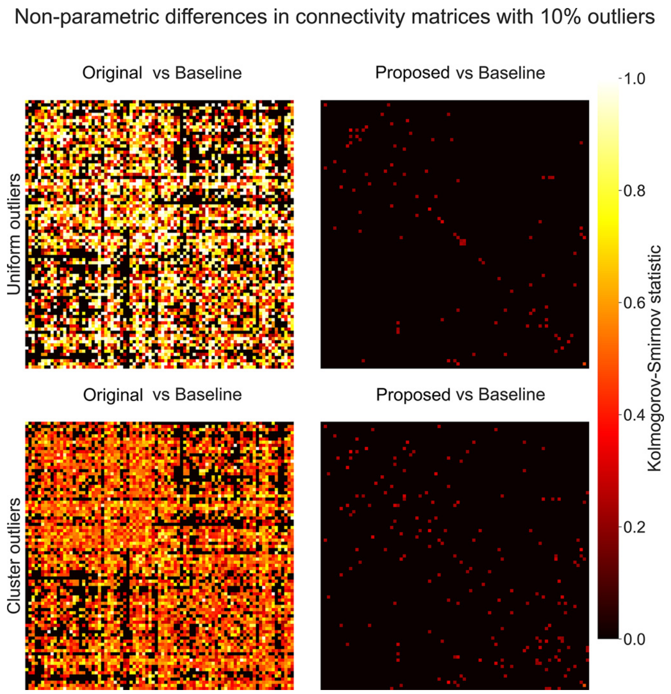

Fig. 9. Differences in the network edges due to outliers measured using non-parametric statistics. The color scale indicates the average effect size calculated as Kolmogorow-Smirnov statistic. The left and right columns show the difference from Baseline to original and proposed robust COMMIT. The top and bottom rows show the results from uniform and clustered outlier schemes. Robust augmentation produces smaller effect sizes which means that the distributions between the Baseline and proposed COMMIT_outlier were very similar in both outlier schemes. For original COMMIT, the uniform outliers produced larger effect sizes than the clustered. This is the outcome of the cumulative distribution function-based statistics where the uniform outliers results in distributions with a high precision but low accuracy which do not overlap with the Baseline whereas the clustered outlier results in distributions with a very low precision but moderate accuracy which do overlap with the Baseline.

To compare our results to literature, we refer to the study by Kunz et al. (2014). Kunz, et al. demonstrated that NODDI-model can be fitted to newborn subjects and they reported average intracellular signal fractions from various brain regions calculated from 13 subjects amongst other results. In their study, the intracellular signal fraction values ranged from 0.19 ± 0.02 in the posterior part of the superior longitudinal fasciculus to 0.45 ± 0.06 in the body of the corpus callosum. While our results depicted in Fig. 10 are vaguely in the same range, the highest observed value (0.35) is lower than in the study by Kunz, et al. This could be due to various reasons ranging from acquisition techniques to the fact that the subject in this study was an extremely preterm born infant and therefore exhibited different anatomical characteristics than the newborns studied by Kunz, et al.

4. Discussion

We demonstrated that tractogram filtering is severely affected by subject motion artefacts and that with our proposed robust augmentation these effects can be mitigated. In clinical research with uncooperative patients such as infants, it is highly likely that motion to some degree occurs during scanning. This leads to corrupted measurements which should not affect any modeling methods applied to the data. To best of our knowledge, this is the first time that motion related outliers are considered in the context of tractogram filtering therefore this update is crucial to enable tractogram filtering in clinical research.

The reason why we evaluated the proposed augmented cost function with simulated brains instead of real brain data was simply to ensure that nothing else in the relatively long dMRI processing pipeline might affect the results. For example, it is currently unknown issue, how outliers affect constrained spherical deconvolution based probabilistic tractographies. While there have been proposals for robust higher order model estimators (Pannek et al., 2012), such are not widely available. Furthermore, developing and evaluation of robustness of currently available constrained spherical deconvolution tractography algorithms are beyond the main scope of this study.

4.1. Comparison to other filtering methods

While similar weighted cost function as in Eq. (1) has been proposed before in SIFT filtering algorithm (Smith et al., 2013), those have been designed and tested to account for partial voluming related artefacts - not subject motion. The main difference in these artefact types is that partial voluming affects all dMRI data whereas subject motion affects only part of the dMRI data randomly. Therefore, adjusting for partial voluming requires one three-dimensional reliability image whereas adjusting for subject motion requires four-dimensional reliability image as the measurement reliability must be accounted for each dMRI data separately. This difference in the implementations of the algorithms also makes the accurate comparison of them fall outside the scope of this study.

4.2. Correcting for artefacts

Our proposed algorithm (Fig. 1 can also be used to adjust for partial voluming but the necessity of that depends on the forward model used in COMMIT. For example, with ball and sticks model, voxels containing cerebrospinal fluid or gray matter can be described with an increased contribution from a ball compartment therefore the contribution of a stick compartment could be correct even without additional reliability weighting. If reliability weights are used, then the estimate for ball compartment would likely be improved but that should not still affect the filtered tractogram.

With motion induced artefacts, the outliers cause anisotropic signal deviations (Sairanen et al., 2017) affecting only part of the dMRI data. Therefore, COMMIT cannot adjust for those deviations simply by increasing the contribution of the ball compartment as the deviations are not isotropic over dMRI measurements. This is demonstrated in Fig. 10 where normal COMMIT obtains incorrect estimates for isotropic signal fraction maps i.e., ball compartments. Issue propagates causing also incorrect estimates for intracellular signal fraction maps i.e., stick contributions. Therefore, a local motion artefact can have a global adverse effect in tractography filtering if not accounted for.

4.3. Statistical analysis

The global connectivity differences (Fig. 5) showed that normal COMMIT results varied heavily depending on the used outlier scheme (uniform or cluster) as well as on the outlier percentage (5% and 10%) whereas the robust COMMIT_outlier results remained relatively intact in all cases. Statistical tests (Table 1) depicted that global connectivity was significantly affected by outliers with normal COMMIT producing also large effect sizes when compared against Baseline. On the contrary, comparison between Baseline and the proposed robust COMMIT_outlier resulted in small effect sizes despite two-sample Kolmogorov-Smirnov tests reporting statistically significant differences in 10% outlier simulations. It is possible that the number of simulated outliers (10%) was already reaching the limit after which the missing data problem becomes too severe even for robust modeling methods. This could also be related to sample size being so large that Kolmogorov-Smirnov test finds any differences statistically significant despite having relatively small effect

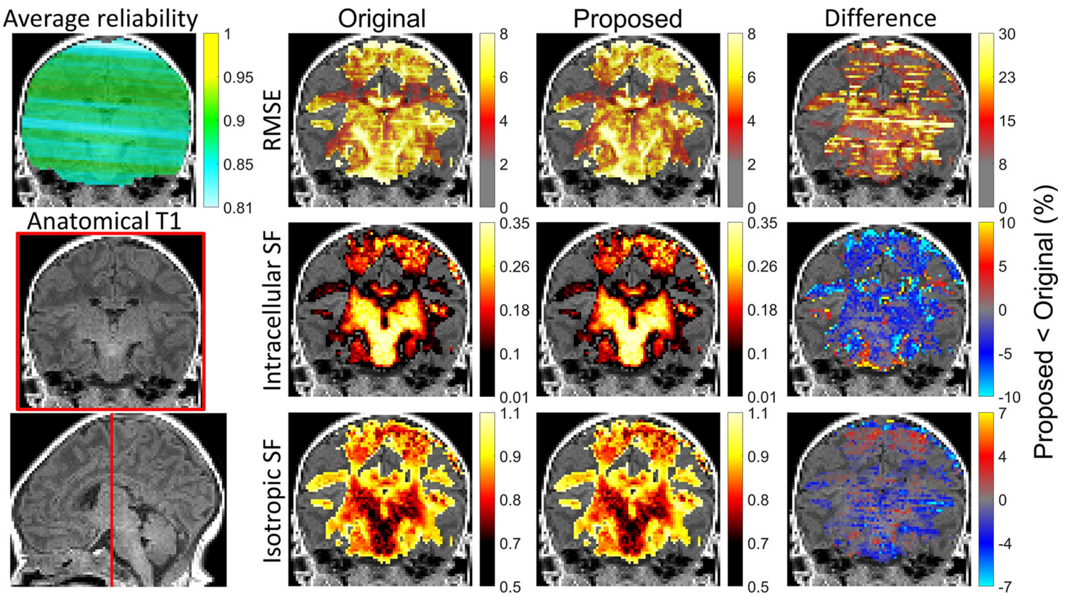

Fig. 10. Summary of the evaluation with in vivo infant data. Coronal images visualize the slicewise artefacts that typically occur in the axial plane. Due to rotational subject motion (yaw, pitch, and roll), the acquired axial plane becomes an oblique plane during the image transformations that are necessary to align images to same spatial coordinates. On the left, an average reliability or confidence map and T1-image visualize the artefactual regions in the measurements after the image alignment. The bottom image in the left column shows a sagittal slice in which the red line indicates the position of the coronal slice. The three columns from the right are results for original COMMIT, proposed robustly weighted COMMIT_outlier and their difference, respectively. The first row details the results for root mean squared error (RMSE), the second row for intracellular signal fractions, and the third row for isotropic signal fractions. In this case, the signal drop outliers are seen as increased diffusivity in random directions and original COMMIT tries to adjust for it by increasing isotropic signal fraction in the affected slices. This results in an increased RMSE in the corresponding slices as well as slightly overestimated isotropic signal fraction which can is easiest to see in the corresponding difference map. This leads to interesting problem elsewhere in the brain (not affected by outliers) where original COMMIT overestimates the intracellular signal fraction. It should be noted that this is a case example which likely cannot be generalized as the effect of outliers is difficult to predict and depends on the affected gradient directions as well as the underlying brain structures.

sizes. Therefore, in future studies some other non-parametric tests could provide better results.

A more in-depth analysis of the connection from medulla to the right precentral gyrus (Figs. 6 and 7) revealed that ANOVA failed to find statistically significant differences between the groups with a p-value of 0.13 in 10% outlier simulations whereas significant differences were found in 5% outlier simulations with p-value less than 0.05. The non-parametric Friedman's tests indicated for both outlier percentages that differences existed between the groups with a p-value less than 0.05. Notably, the effect sizes in comparison between Baseline and robust COMMIT_outlier remained much smaller than in comparisons between Baseline and normal COMMIT providing support for our proposed method being capable to mitigate these artefacts even for individual network edges.

In summary, it remains unsolved what test statistic would be the most suitable to analyze such data that is affected by outliers in anisotropic manner. We used two alternative approaches to evaluate the differences in group averages (ANOVA) and group distributions (Kolmogorov-Smirnov). Average based analyses are likely inefficient to locate all differences arising from outliers in the data whereas non-parametric tests can be even too sensitive to baseline shifts. Therefore, instead of statistical significance, the obtained effect sizes are likely more meaningful results.

4.4. Robust modeling vs outlier replacement

This section extends outside the main scope of this study and is intended for the readers interested in slicewise outliers and how they should be addressed in dMRI in general. We added this section because we feel that the use of outlier replacement in diffusion weighted literature is not truly justified and should not be continued in its current state. To understand our reasoning, readers are encouraged to familiarize the concept of outlier replacement which in statistics is known as data imputation. For this, we recommend the textbook *Statistical analysis with missing data* by Little (2002).

Outlier replacement in dMRI is a form of multiple imputation which has been developed to correct for missing data in statistical analyses. Benefits of well performed imputation include decreased bias, increased precision, and most conveniently the ability to apply standard statistical tests and model estimators. For example, applying the standard *t*-test on sample that contains many missing or incorrect measurements could result in highly incorrect outcome whereas using a fixed sample that contains correctly imputed data could provide more reasonable results. This is, of course, the reason why outlier replacement seems so tempting in the context of dMRI: simply replace outliers and use the rest of the analysis pipeline as it is.

Imputation methods range from naive neighborhood interpolation (outlier is replaced e.g., by the average of its neighbors) and sample statistic-based replacements (outlier is replaced by e.g., the mean or median value of the sample) to complex model prediction-based replacements. Some of these ideas have already been transferred to dMRI usage by replacing outliers by their q-space neighborhood (Niethammer et al., 2007) or model-based estimations (Lauzon et al., 2013; Andersson et al., 2016; Koch et al., 2019). These methods have in common that they depend on perfect outlier detection which can be problematic if there are many outliers. If some of the outliers are not correctly detected, this can lead to bias in the interpolation or modeling used in imputation which would propagate to the diffusion modeling that is performed using a normal estimator.

The idea in robust modeling is to account for the unreliability of the measurements and weigh each data point accordingly. In dMRI, such reliability can be obtained from voxelwise residuals (outliers tend to

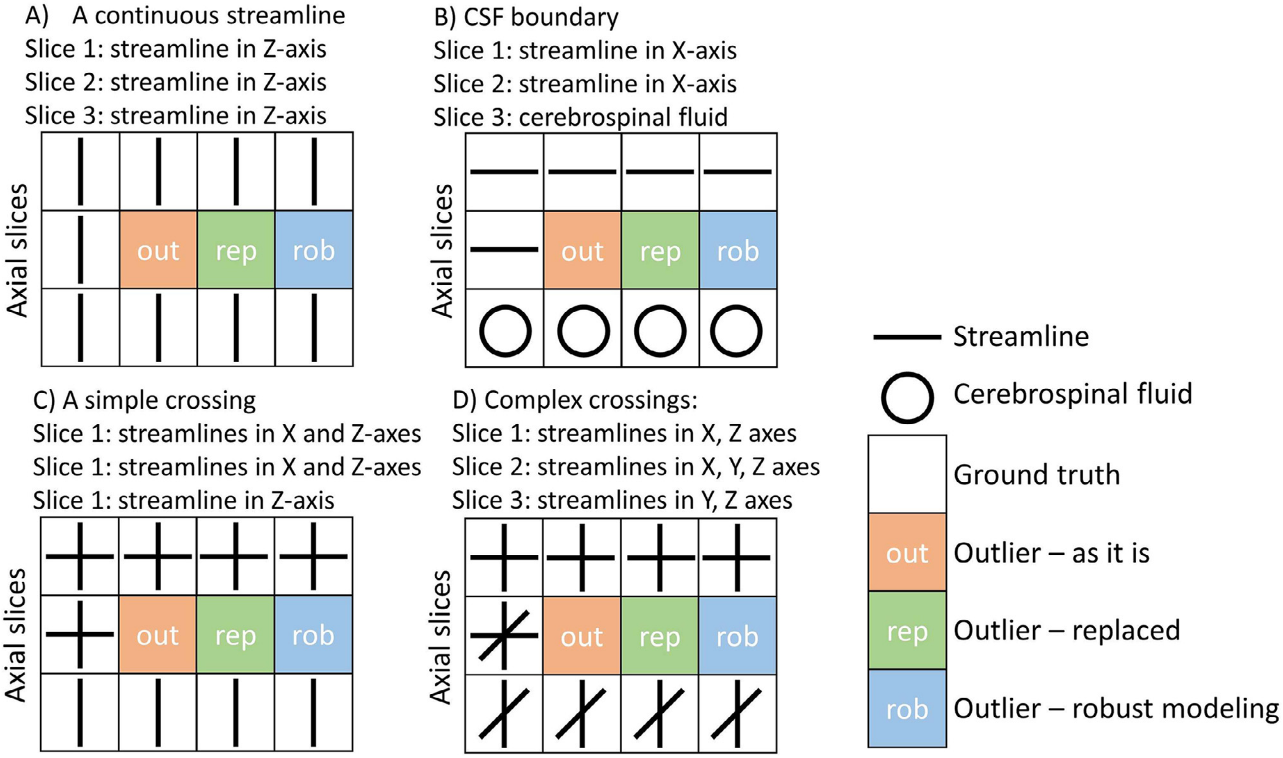
A) A continuous streamline
Slice 1: streamline in Z-axis
Slice 2: streamline in Z-axis
Slice 3: streamline in Z-axis
B) CSF boundary
Slice 1: streamline in X-axis
Slice 2: streamline in X-axis
Slice 3: cerebrospinal fluid
C) A simple crossing
Slice 1: streamlines in X and Z-axes
Slice 1: streamlines in X and Z-axes
Slice 1: streamline in Z-axis
D) Complex crossings:
Slice 1: streamlines in X, Z axes
Slice 2: streamlines in X, Y, Z axes
Slice 3: streamline in Y, Z axes
Axial slices
out
rep
rob
Streamline
Cerebrospinal fluid
Ground truth
Outlier – as it is
Outlier – replaced
Outlier – robust modeling

Fig. 11. Illustration of the streamline phantom setups used to investigate differences between outlier replacement and robust modeling. The left column in all setups is a ground truth with no outlier and normal modeling, the second column is a control with outlier and normal modeling, the third column is replaced outlier with normal modeling, and the fourth column is outlier and robust modeling. Outlier replacement was calculated as an average from the slice below and above the outlier (i.e., spatial neighborhood). (A) A single streamline traverses in parallel to z-axis throughout the phantom. Therefore, the outlier replacement produces exactly the missing data and should provide the same result with the ground truth. (B) A single streamline traverses parallel to x-axis in two upper slices with the third slice consisting of cerebrospinal fluid. (C) Two upper slices consist of streamline crossings in x and z-axes with the third consisting of streamline traversing only parallel to z-axis. (D) The first slice consists of streamline crossings in x and z-axes, the middle slice consists of crossings in all main axes, and the third slice in crossing in y and z-axes.

have large residuals) which has been implemented in algorithms suchs as RESTORE (Chang et al., 2005; Chang et al., 2012) and REKINDLE (Tax et al., 2015) for tensor model fitting. Similar approach can be applied to nearly any model.

Voxelwise outlier detection, however, is suboptimal in dMRI because artefacts in the echo-planar imaging result in whole slices being incorrect. Therefore, detecting subject motion related outliers in slicewise manner and assigning the reliability to all voxels in those slices is arguably more powerful approach (Andersson et al., 2016; Sairanen et al., 2018). Moreover, slicewise outliers can be used as complementary information for voxelwise estimators to adjust for more local sources of uncertainties e.g., pulsation due to heartbeat.

To illustrate the performance of the aforementioned ideas, we provide a minimal example in which we compare naive outlier replacement to simple robust model estimation. We used the constrained spherical deconvolution (CSD) algorithm implemented in DIPY (Tournier et al., 2007; Garyfallidis et al., 2014) in this evaluation as evaluating the differences using COMMIT would be computationally inefficient (and well beyond the purpose of the current study).

In CSD, we used the default DIPY-library parameters with Lmax 8, tau 0.1, 362 vertices on the symmetrical sphere, relative peak threshold of 0.5, minimal peak difference angle of 25°, and 50 iterations. We developed four streamline setups that are described in Fig. 11. All setups consisted of three axial slices in which the middle slice was affected by a full signal dropout artefact. We investigated what happens to CSD signal prediction if *i*) nothing was done to the outlier, *ii*) outlier was replaced with a naive neighborhood interpolation, and *iii*) spherical harmonic coefficients used in CSD were obtained using a robust in-house version of CSD algorithm. The in-house algorithm simply decreased the outlier weight to zero in the linear least squares estimation of the spherical harmonic coefficients.

We used infinite signal-to-noise ratio to evaluate only the effects of the signal dropout. Outliers were introduced incrementally from 0 to 9 of one of the HCP gradient scheme shells with b-value of $2000 s/mm^2$using three different schemes in outlier selection. The first scheme represented the worst possible situation where outliers were clustered in the *q*-space (e.g. Fig. 4), the second scheme represented the best possible situation where outliers were uniformly placed in the *q*-space based on their electrostatic repulsion (Sairanen et al., 2017), and the third scheme represented randomly selected outliers. In the first two cases, we evaluated all possible 90 cases whereas the random scheme consisted of 500 combinations. Signal predictions from these schemes were compared to prediction from the ground truth fit that was not affected by outliers. In this simulation we did not add Rician noise, therefore direct comparison to ground truth is sound.

It should be noted that random outlier scheme is not the perfect method to evaluate these effects in dMRI model estimation due to the extremely large number of possible combinations. For example, selecting 9 outliers out of 90 diffusion weighted images can be performed in $7.0625 \cdot 10^{11}$different ways therefore selecting 500 of these randomly might not represent the population well. Since evaluating all possible combinations is possible only for a smaller number of gradient directions (Sairanen et al., 2017), it is more informative to investigate the best (uniform) and the worst (clusters) extreme cases in the context of model estimator algorithm development. Practically having a subject that has either of the worst or the best-case outlier setup is very small. Therefore, if the interest is not the behavior of an algorithm but a dis-

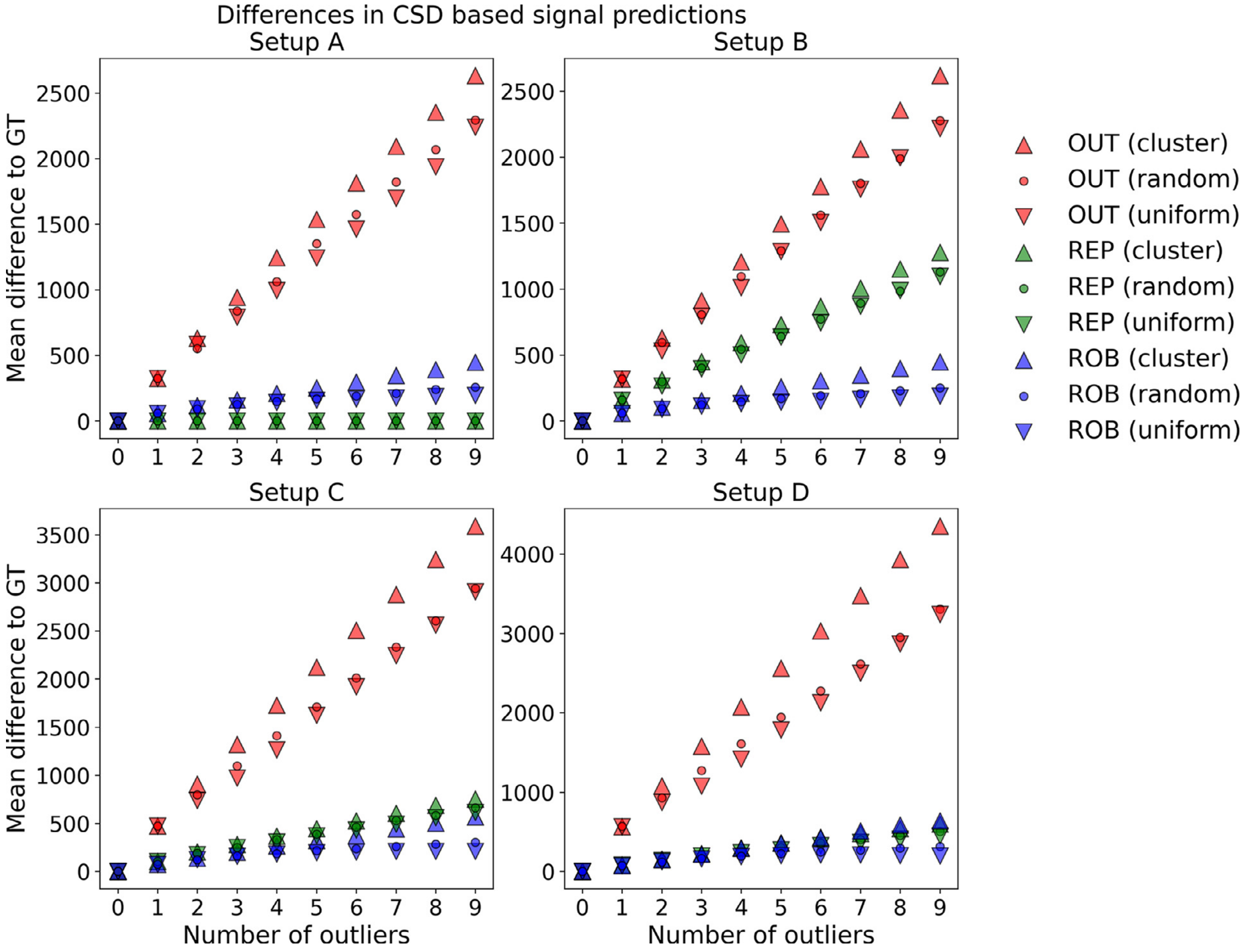

Fig. 12. Outlier replacement compared to robust modeling in constrained spherical deconvolution (CSD). Upward triangles (▲) indicate clustered outliers, dots (○) indicate randomly placed outliers, and downward triangles (▼) indicate uniformly placed outliers. All values are calculated as an average residual from ground truth (GT) signal prediction. Red markers (OUT) are results from normal CSD, green markers (REP) are results from normal CSD with outlier replacement, and blue markers (ROB) are results from the robust CSD. The clustered outliers (▲) result in the largest differences in all cases and the uniform outliers (▼) result in the smallest differences. Random cases (○) are generally closer to the uniform situation as the extreme outliers tend to result in heavily tailed distributions. Outlier replacement provided better results only in the Setup A in which the outlier could be replaced with similar data that was missing. In all other cases, where the spatial neighborhood did not exactly represent the missing measurement, the outlier replacement produced larger difference to the ground truth than the robust estimator. As the complexity of the streamline phantom increases in Setups C and D, the difference between replacement and robust methods become smaller. Importantly, both the robust modeling and outlier replacement improved the CSD prediction compared to the baseline case (OUT).

tribution of values such as FA, it can be more informative to investigate random cases than either of these extremes.

Results of these four setups, shown in Fig. 12 demonstrate that if nothing is done to the outliers, the difference to ground truth increases linearly as the number of outliers increase. Of course, after the mathematical problem becomes ill-conditioned a more chaotic results would be expected (Sairanen et al., 2017) but this is likely to occur with much larger number of outliers. Based on the Fig. 12 the naive outlier replacement outperformed robust modeling only in the Setup A in which the outlier was replaced by identical information from the neighboring voxels. In setups B, C, and D with more complex and perhaps realistic streamline combinations robust modeling outperformed outlier replacement by providing results that were closer to the ground truth.

It can be reasonably argued that the rather naive outlier replacement we implemented here could be improved with already available proposals of *q*-space neighborhood (Niethammer et al., 2007) or model-based estimations (Lauzon et al., 2013; Andersson et al., 2016; Koch et al., 2019). However, same applies to the in-house robust spherical harmonic linear least squares estimator we developed for this task which simply down weighs the outlier measurements. However, while it might be possible to fine tune outlier replacement in dMRI to the degree that matches the robust modeling, outlier replacement would still lack the ability to evaluate the uncertainty in the fitted model rendering it less useful method for clinical usage that might require or benefit from knowledge of the method's uncertainty (e.g., surgery or radiotherapy).

The reader might be confused by the previous statement that imputation could not contain information about uncertainties while even the text book by Little (2002) we cited has a chapter called "*Estimation of Imputation Uncertainty*". To understand this, remember that the imputed sample in dMRI is generally an axial slice in a three-dimensional stack of slices that are a part of four-dimensional series of diffusion weighted images. MRI scan of the whole series takes several minutes during which the patient's head tends to move and especially rotate (yaw, pitch, and roll). These rotated images must be aligned with some reference image before model fitting but by doing so the image registration transforms the axial (outlier) slice into an oblique plane which is an interpolation between the slice and its neighbors.

This process of image alignment is described in Fig. 1 of Sairanen et al. (2018) but in short, afterwards it is likely impossible to accurately distinguish an imputed signal fraction from a normal signal. This means that likelihood-based estimators or bootstrap methods

(Whitcher et al., 2008) no longer can estimate the uncertainty of the fitted diffusion model. On the contrary, robust modeling that is based on measurement reliability weights would still be able to tell this difference. Therefore, any clinical application that might benefit of these uncertainty estimates would be hindered by using outlier replacement. To avoid such bottleneck in the future of dMRI, we argue that it would be highly beneficial for the dMRI community to avoid using the outlier replacement in its current form. Even in basic neuroscience, it could be beneficial to know the voxelwise distributions of model derived values such as fractional anisotropy (e.g., $FA = 0.6 \pm 0.03$) to perform sound statistical analyses.

4.5. Where to go from here?

We considered only post-scan motion corrections in this study because during-scan corrections should be able to produce data that does not need these correction algorithms. The problem with during-scan corrections is their limited availability due to external hardware requirements or still experimental software. Due to the long-time span of tens of years required to advance MRI technology in clinical use, it is unlikely that these during-scan correction methods would be so widely available in clinical research centers that post-scan corrections such as our proposal are rendered obsolete any time soon. While the post-scan corrections are more like a remedy to the symptom instead of cure to the cause, novel studies on clinical patients and even infants are increasingly proposed and carried out therefore the need for robust tools is current and cannot wait decades for hardware-based solutions.

While we discussed using this robust weighing to adjust for subject motion related signal dropout artefacts in this study, it could be possible to use this same approach to correct for signal dropouts that have different origin. For example, in preclinical in-vivo and ex-vivo settings acquiring very thin slices can be very demanding for the scanner gradient system. This can cause gradient system malfunctions that result in similar signal dropouts as the subject motion (Le Bihan, et al. 2006). In such case, the proposed approach of outlier detection followed by robust modeling could be beneficial.

Conclusion

We proposed an augmentation to the tractogram algorithm COMMIT that renders it robust towards subject motion outliers in the measurements. This addition is necessary for conducting tractogram filtering in clinical research where subject motion is often unavoidable. While robust data processing has been implemented before in the context of diffusion tensor and higher order model estimations, it has not been previously implemented for tractogram filtering. We used realistic whole brain Monte-Carlo simulations that account for kissing and crossing fiber structure as well as partial volumeing to successfully demonstrate that our augmentation is capable to accurately map the structural brain connectivity in the presence of such outliers in the data. We also demonstrated that if this correction is not done, the structural connectivity estimates can become strongly biased. With this update any clinical study investigating structural connectomics of children or uncooperative patient populations can robustly perform their analyses without the need to exclude subjects with outliers from them.

Credit authorship contribution statement

Viljami Sairanen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Mario Ocampo-Pineda:** Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing. **Cristina Granziera:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Funding acquisition. **Simona Schiavi:** Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Alessandro Daducci:** Conceptualization, Methodology, Project administration, Software, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgements

V.S. was supported by the Brain Research Foundation Verona for the period devoted to the development of the method and the drafting of the manuscript, as well as by the Orion Research Foundation sr and Instrumentarium Science Foundation sr. during data analysis, visualization, manuscript finalization, and revision process. M.O-P., S.S. and A.D have no relevant financial or non-financial interests to disclose. C.G. was funded by the Swiss National Science Foundation (SNSF) grant PP00P3_176984, the Stiftung zur Förderung der gastroenterologischen und allgemeinen klinischen Forschung, EUROSTAR E!113682 HORIZON2020. The authors wish to thank the Finnish Grid and Cloud Infrastructure (FGCI) for supporting this project with computational and data storage resources. Open access funded by Helsinki University Library

Data and code availability statement

Data used in the study entitled "**Incorporating outlier information into diffusion MR tractogram filtering for robust structural brain connectivity and microstructural analyses**" has two sources: The first source is freely available data from the Human Connectome Project subject specified in the manuscript that is used in the evaluation simulations. This data is shared via the Human Connectome Project. The second source is data from an ongoing clinical study which ethical agreement sets tight restrictions on who can access the data and for what purposes. Therefore, the second data cannot be freely shared. To access this data, it is necessary to obtain a collaboration agreement with the clinical site as well as to update the ethical approval.

Software code used in the paper is distributed freely via Github repository https://github.com/daducci/COMMIT.

References

Alexander, D.C., Dyrby, T.B., Nilsson, M., Zhang, H., 2019. Imaging brain microstructure with diffusion MRI: practicality and applications. NMR Biomed. 32 (4). doi:10.1002/nbm.3841.

Andersson, J.L.R., Graham, M.S., Zsoldos, E., Sotiropoulos, S.N., 2016. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. Neuroimage 141, 556–572. doi:10.1016/j.neuroimage.2016.06.058.

Basser, P.J., Mattiello, J., LeBihan, D., 1994. MR Diffusion Tensor Spectroscopy and Imaging. Biophys. J. 66 (1), 259–267. doi:10.1016/S0006-3495(94)80775-1.

Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. *In vivo* fiber tractography using DT-MRI data. Magn. Reson. Med. 44 (4), 625–632. doi:10.1002/1522-2594(200010)44:4-625::AID-MRM17-3.0.CO;2-O.

Benitez, A., Fieremans, E., Jensen, J.H., Falangola, M.F., Tabesh, A., Ferris, S.H., Helpern, J.A., 2014. White matter tract integrity metrics reflect the vulnerability of late-myelinating tracts in Alzheimer's Disease. NeuroImage Clin. 4 (January), 64–71. doi:10.1016/j.nicl.2013.11.001.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodol.) 57 (1), 289–300.

Chang, L.C., Jones, D.K., Pierpaoli, C., 2005. RESTORE: robust estimation of tensors by outlier rejection. Magn. Reson. Med. 53 (5), 1088–1095. doi:10.1002/mrm.20426.

Chang, L.C., Walker, L., Pierpaoli, C., 2012. Informed RESTORE: a method for robust estimation of diffusion tensor from low redundancy datasets in the presence of physiological noise artifacts. Magn. Reson. Med. 68 (5), 1654–1663. doi:10.1002/mrm.24173.

Daducci, A., Palù, A.D., Lemkaddem, A., Thiran, J.P., 2015. COMMIT: convex optimization modeling for microstructure informed tractography. IEEE Trans. Med. Imaging 34 (1), 246–257. doi:10.1109/TMI.2014.2352414.

Delettre, C., Messé, A., Dell, L.A., Foubet, O., Heuer, K., Larrat, B., Meriaux, S., et al., 2019. Comparison between diffusion MRI tractography and histological tract-tracing of cortico-cortical structural connectivity in the ferret brain. Netw. Neurosci. 3 (4), 1038–1050. doi:10.1162/netn_a_00098.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into Gyral based regions of Interest. Neuroimage 31 (3), 968–980. doi:10.1016/j.neuroimage.2006.01.021.

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

Fieremans, E., Benitez, A., Jensen, J.H., Falangola, M.F., Tabesh, A., Deardorff, R.L., Spampinato, M.V.S., et al., 2013. Novel white matter tract integrity metrics sensitive to Alzheimer disease progression. Am. J. Neuroradiol. 34 (11), 2105–2112. doi:10.3174/ajnr.A3553.

Fischl, B., 2012. FreeSurfer. NeuroImage, 20 YEARS OF fMRI 62 (2), 774–781. doi:10.1016/j.neuroimage.2012.01.021.

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

Genc, S., Malpas, C.B., Holland, S.K., Beare, R., Silk, TJ., 2017. Neurite density index is sensitive to age related differences in the developing brain. Neuroimage 148, 373–380. doi:10.1016/j.neuroimage.2017.01.023, March.

Griffa, A., Baumann, P.S., Thiran, J.P., Hagmann, P., 2013. Structural Connectomics in Brain Diseases. NeuroImage Mapp Connect. 80, 515–526. doi:10.1016/j.neuroimage.2013.04.056, October.

Gudbjartsson, H.Á.K., Patz, S., 1995. The rician distribution of noisy MRI data. Magn. Reson. Med. 34 (6), 910–914. doi:10.1002/mrm.1910340618.

Horsfield, M.A., Jones, D.K., 2002. Applications of diffusion-weighted and diffusion tensor MRI to white matter diseases-a review. NMR Biomed. 15 (7–8), 570–577. doi:10.1002/nbm.787.

Huber, E., Neto Henriques, R., Owen, J.P., Rokem, A., Yeatman, J.D., 2019. Applying microstructural models to understand the role of white matter in cognitive development. Dev. Cogn. Neurosci. 36 (April), 100624. doi:10.1016/j.dcn.2019.100624.

Jeurissen, B., Tournier, J.-.D., Dhollander, T., Connelly, A., Sijbers, J., 2014. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. Neuroimage 103, 411–426. doi:10.1016/j.neuroimage.2014.07.061, December.

Kamiya, K., Hori, M., Aoki, S., 2020. NODDI in clinical research. J. Neurosci. Methods 346, 108908. doi:10.1016/j.jneumeth.2020.108908, December.

Koch, A., Zhukov, A., Stöcker, T., Groeschel, S., Schultz, T., 2019. SHORE-based detection and imputation of dropout in diffusion MRI. Magn. Reson. Med. 82 (6), 2286–2298. doi:10.1002/mrm.27893.

Kunz, N., Zhang, H., Vasung, L., O'Brien, K.R., Assaf, Y., Lazeyras, F., Alexander, DC., Hüppi, PS., 2014. Assessing white matter microstructure of the newborn with multishell diffusion MRI and biophysical compartment models. Neuroimage 96, 288–299. doi:10.1016/j.neuroimage.2014.03.057, August.

Lauzon, C.B., Asman, A.J., Esparza, M.L., Burns, S.S., Fan, Q., Gao, Y., Anderson, A.W., Davis, N., Cutting, L.E., Landman, B.A., 2013. Simultaneous analysis and quality assurance for diffusion tensor imaging. PLoS ONE 8 (4). doi:10.1371/journal.pone.0061737.

Le Bihan, D., Poupon, C., Amadon, A., Lethimonnier, F., 2006. Artifacts and pitfalls in diffusion MRI. J. Magn. Reson. Imaging 24 (3), 478–488 An Official Journal of the International Society for Magnetic Resonance in Medicine.

Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. In: Proceedings of the 17th Scientific Meeting, International Society for Magnetic Resonance in Medicine, 17, p. 3537 –3537.

Little, R.J.A., 2002. Statistical analysis with missing data. Statistical Analysis with Missing Data. John Wiley & Sons, Inc, Hoboken, New Jersey.

Maier-Hein, K.H., Neher, P.F., Houde, J.C., Côté, M.A., Garyfallidis, E., Zhong, J., Chamberland, M., et al., 2017. The challenge of mapping the human connectome based on diffusion tractography. Nat. Commun. 8 (1), 1–13. doi:10.1038/s41467-017-01285-x.

Niethammer, M., Bouix, S., Aja-Fernández, S., Westin, C.F., Shenton, M.E., 2007. Outlier rejection for diffusion weighted imaging. Med. Image Comput. Comput. Assist. Interv. 10 (Pt 1), 161–168.

Novikov, D.S., Fieremans, E., Jespersen, S.N., Kiselev, V.G., 2019. Quantifying brain microstructure with diffusion MRI: theory and parameter estimation. NMR Biomed. 32 (4), e3998. doi:10.1002/nbm.3998.

Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., Johnson, H.J., Styner, M., 2014. DTIPrep: quality control of diffusion-weighted images. Front. Neuroinform. 8, 4. doi:10.3389/fninf.2014.00004, –4.

Panagiotaki, E., Schneider, T., Siow, B., Hall, M.G., Lythgoe, M.F., Alexander, D.C., 2012. Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison. Neuroimage 59 (3), 2241–2254. doi:10.1016/j.neuroimage.2011.09.081.

Pannek, K., Fripp, J., George, J.M., Fiori, S., Colditz, P.B., Boyd, R.N., Rose, S.E., 2018. Fixel-based analysis reveals alterations is brain microstructure and macrostructure of preterm-born infants at term equivalent age. NeuroImage Clin. 18 (January), 51–59. doi:10.1016/j.nicl.2018.01.003.

Pannek, K., Raffelt, D., Bell, C., Mathias, J.L., Rose, S.E., 2012. HOMOR: higher order model outlier rejection for high b-value MR diffusion data. Neuroimage 63 (2), 835–842. doi:10.1016/j.neuroimage.2012.07.022.

Pecheva, D., Tournier, J.D., Pietsch, M., Christiaens, D., Batalle, D., Alexander, D.C., Hajnal, J.V., Edwards, A.D, Zhang, H., Counsell, S.J., 2019. Fixel-based analysis of the preterm brain: disentangling bundle-specific white matter microstructural and macrostructural changes in relation to clinical risk factors. NeuroImage Clin. 23, 101820. doi:10.1016/j.nicl.2019.101820, January.

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. doi:10.1016/j.neuroimage.2015.06.068, October.

Sairanen, V., Kuusela, L., Sipilä, O., Savolainen, S., Vanhatalo, S., 2017. A novel measure of reliability in diffusion tensor imaging after data rejections due to subject motion. Neuroimage 147. doi:10.1016/j.neuroimage.2016.11.061.

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. doi:10.1016/j.neuroimage.2018.07.003, November.

Sairanen, V., Ocampo-Pineda, M., Granziera, C., Schiavi, S., Daducci, A., 2021. Enhancing reliability of structural brain connectivity with outlier adjusted tractogram filtering. In: Proceedings of the IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE.

Samani, Z.R., Alappatt, J.A., Parker, D., Ould Ismail, A.A., Verma, R., 2019. QC-Automator: deep learning-based automated quality control for diffusion MR images. Front. Neurosci. 13, 1456. doi:10.3389/fnins.2019.01456.

Schiavi, S., Petracca, M., Battocchio, M., El Mendili, MM., Paduri, S., Fleysher, L., Inglese, M., Daducci, A., 2020. Sensory-motor network topology in multiple sclerosis: structural connectivity analysis accounting for intrinsic density discrepancy. Hum. Brain Mapp. 41 (11), 2951–2963. doi:10.1002/hbm.24989.

Sheskin, D.J., 2004. Handbook of Parametric and Nonparametric Statistical Procedures, 2nd Ed. Chapman & hall/CRC, Boca Raton, FL.

Smith, RE., Tournier, J.D., Calamante, F., Connelly, A., 2012. Anatomically-constrained tractography: improved diffusion MRI streamlines tractography through effective use of anatomical information. Neuroimage 62 (3), 1924–1938. doi:10.1016/j.neuroimage.2012.06.005.

Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A., 2013. SIFT: spherical-deconvolution informed filtering of tractograms. Neuroimage 67, 298–312.

Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A., 2015. SIFT2: enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. Neuroimage 119, 338–351.

Tax, C.M.W., Otte, W.M., Viergever, M.A., Dijkhuizen, RM., Leemans, A., 2015. REKINDLE: robust extraction of kurtosis INDices with linear estimation. Magn. Reson. Med. 73 (2), 794–808. doi:10.1002/mrm.25165.

Thomas, C., Ye, F.Q., Okan Irfanoglu, M., Modi, P., Saleem, KS., Leopold, D.A., Pierpaoli, C., 2014. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. Proc. Natl. Acad. Sci. 111 (46), 16574–16579. doi:10.1073/pnas.1405672111.

Tournier, J.D, Calamante, F., Connelly, A., 2007. Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. Neuroimage 35 (4), 1459–1472. doi:10.1016/j.neuroimage.2007.02.016.

Tournier, J.D, Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.H., Connelly, A., 2019. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. Neuroimage 202, 116137. doi:10.1016/j.neuroimage.2019.116137, November–116137.

Tournier, J.D., Calamante F., and Connelly A.. 2010. "Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions," 1.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-minn human connectome project: an overview. Neuroimage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041, October.

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. doi:10.1016/j.neuroimage.2013.05.028.

Vos, S.B., Tax C.M.W., Luijten P.R., Ourselin S., Leemans A., and Froeling M.. 2016 "The importance of correcting for signal drift in diffusion MRI." 10.1002/mrm.26124.

Whitcher, B., Tuch, D.S., Wisco, J.J., Gregory Sorensen, A., Wang, L., 2008. Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. Hum. Brain Mapp. 29 (3), 346–362. doi:10.1002/hbm.20395.

Yeh, C.H., Jones, D.K., Liang, X., Descoteaux, M., Connelly, A., 2020. Mapping Structural Connectivity Using Diffusion MRI: challenges and Opportunities. J. Magn. Reson. Imaging 1–17. doi:10.1002/jmri.27188.

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

Zhang, F., Daducci A., He Y., Schiavi S., Seguin C., Smith R., Yeh C.H., Zhao T., and O'Donnell L.J.. 2021. "Quantitative mapping of the brain's structural connectivity using diffusion mri tractography: a review." ArXiv:2104.11644 [q-Bio], April. http://arxiv.org/abs/2104.11644.

Zöllei, L., Iglesias, J.E., Ou, Y., Grant, P.E, Fischl, B., 2020. Infant FreeSurfer: an automated segmentation and surface extraction pipeline for T1-weighted neuroimaging data of infants 0–2 years. Neuroimage 218, 116946. doi:10.1016/j.neuroimage.2020.116946, September.