Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images

Jesper L. R. Andersson[1*], Mark S. Graham[**], Enikő Zsoldos[***], and Stamatios N. Sotiropoulos[*]

[*]FMRIB Centre, Oxford University, Oxford, United Kingdom

[**]Centre for Medical Image Computing & Department of Computer Science, University College London, London, United Kingdom

[***]Department of Psychiatry, Oxford University, Oxford, United Kingdom

May 25, 2016

[1]Corresponding author

FMRIB Centre

JR Hospital

Headington

Oxford OX3 9DU

phone: 44 1865 222 782

fax: 44 1865 222 717

mail: jesper.andersson@ndcn.ox.ac.uk

1 ABSTRACT

Despite its great potential in studying brain anatomy and structure, diffusion magnetic resonance imaging (dMRI) is marred by artefacts more than any other commonly used MRI technique. In this paper we present a non-parametric framework for detecting and correcting dMRI outliers (signal loss) caused by subject motion.

Signal loss (dropout) affecting a whole slice, or a large connected region of a slice, is frequently observed in diffusion weighted images, leading to a set of unusable measurements. This is caused by bulk (subject or physiological) motion during the diffusion encoding part of the imaging sequence. We suggest a method to detect slices affected by signal loss and replace them by a non-parametric prediction, in order to minimise their impact on subsequent analysis. The outlier detection and replacement, as well as correction of other dMRI distortions (susceptibility-induced distortions, eddy currents (EC) and subject motion) are performed within a single framework, allowing the use of an integrated approach for distortion correction. Highly realistic simulations have been used to evaluate the method with respect to its ability to detect outliers (type 1 and 2 errors), the impact of outliers on retrospective correction of movement and distortion and the impact on estiation of commonly used diffusion tensor metrics, such as fractional anisotropy (FA) and mean diffusivity (MD). Data from a large imaging project studying older adults (the Whitehall Imaging sub-study) was used to demonstrate the utility of the method when applied to datasets with severe subject movement.

The results indicate high sensitivity and specificity for detecting outliers and that their deleterious effects on FA and MD can be almost completely corrected.

1.1 Keywords

Diffusion, movement, signal loss, outlier, registration.

2 INTRODUCTION

The advent of MR diffusion imaging has led to great progress in our understanding of the microstructure of the brain (see for example Jones (2011) or Johansen-Berg and Behrens (2014) for recent summaries of the field). In the last few years, advances in acquisition (Moeller et al. (2010), Setsompop et al. (2012), Heidemann et al. (2012), Uğurbil et al. (2013), Vu et al. (2015)) have constituted a "second technical revolution" that has enabled acquisition of data of previously unimaginable quality. Recent projects have seen the acquisition and public dissemination of datasets with high spatial resolution, hundreds of diffusion directions and multiple shells (Uğurbil et al. (2013), Sotiropoulos et al. (2013) and (Setsompop et al., 2013)).

However, great care must be taken with the processing of diffusion images since they are marred by artefacts more than any other commonly used MR technique (Andersson and Skare (2011) and Pierpaoli (2011)). In addition, some of the recent advances have come at the cost of increased artefacts in the data straight off the scanner. For example high fields (Vu et al. (2015)) will increase the susceptibility induced off-resonance field. There is a trade-off between in-plane acceleration (IPAT) and simultaneous multi-slice acquisition (MB), and some have decided against IPAT for that reason (Uğurbil et al. (2013)). SNR considerations may cause one to opt for Stejskal-Tanner diffusion encoding in lieu of an EC-compensated alternative (Uğurbil et al. (2013)).

Artefacts in diffusion imaging include

- Image distortions due to field inhomogeneities caused by the subject's head (susceptibility).
- Image distortions due to eddy currents caused by rapid gradient switching during the diffusion encoding.
- Location (of the object within the FOV) changes caused by subject movement.

- Signal dropout caused by bulk motion (subject movement or cardiac pulsation) during the diffusion encoding part of the imaging sequence.

In a recent paper we reported on progress with correcting for the first three of these problems (Andersson and Sotiropoulos (2016)). The current paper deals with the fourth; the signal dropout caused by bulk motion.

Signal dropout is a potentially serious problem that can cause false positives when comparing groups where subjects in one group are more likely to have moved or when regressing data against a factor that is correlated with the degree of subject movement (Yendiki et al., 2014). Examples of when this might occur include the comparison of Parkinson's disease patients with healthy controls (Perea et al. (2013)), regressing against severity of tremor, and developmental studies of babies and young children (Walker et al. (2016)) where motion and dropouts can be severe. In general, the manifestation of the problem may be less obvious, yet may cause biased estimates due to the presence of outlier measurements (Chang et al. (2005)).

Strategies for dealing with signal dropout have received less attention than the related problem of EC-induced distortions. When bulk motion is caused by cardiac induced pulsatile movement, gating has been suggested to avoid the part of the cardiac cycle with the greatest motion (Skare and Andersson (2001) and Nunes et al. (2005)). However, this comes at the cost of a more complicated setup, sequence programming and slower acquisition. Moreover, it has been shown that gating can sometimes be deleterious (Mohammadi et al. (2013)).

Another strategy has been to identify affected voxels (Chang et al. (2005), Chang et al. (2012), Pannek et al. (2012), Collier et al. (2015) and Tax et al. (2015)), "patches" (Zwiers (2010)), slices (Zhou et al. (2011)) or volumes (Oguz et al. (2014)) and to exclude these from further processing. One can also use an estimator that is less sensitive to extreme values (Mangin et al. (2002)) to reduce the effects of outliers. A potential problem with these voxel-wise

approaches is finding a balance between sensitivity and specificity since the SNR for a single voxel can be quite low, especially for deep brain voxels in high b-value data. They are also sensitive to the choice of model for the diffusion signal since the outlier is really defined as the deviation from the model fit. If the model, such as the diffusion tensor (Chang et al. (2005), Chang et al. (2012), Zwiers (2010), Collier et al. (2015) and Tax et al. (2015)), is not able to fully capture the variability of the true signal, a perfectly good voxel may appear to be an outlier.

Another question is the order in which to perform motion correction and outlier detection. It can be problematic to perform motion correction in the presence of outliers since the dropout is likely to affect the estimation of movement. On the other hand, it is not really meaningful to perform a model-to-observation comparison before movement has been properly corrected. Therefore, instead of a model driven approach, Zhou et al. (2011) used local texture analysis to find outlier slices and Oguz et al. (2014) used correlation between adjacent slices in the same volume.

Below we present a novel approach for detecting outliers that are due to signal dropouts. Compared to previous methods, the proposed technique

- Does not rely on a specific biophysical model of the diffusion process to detect outliers.
- Is able to replace outliers with non-parametric signal predictions made at any direction and b-value using angularly neighbouring measurements.
- Is extensively validated against "ground truth" obtained from a recent highly realistic simulator (Graham et al. (2016)).
- Detects outliers in scan space, thereby avoiding the problem of interpolation that mixes valid and outlier measurements.
- Avoids the problem of order of movement/distortion correction and outlier detection.

This novel approach is integrated into our previous framework for joint correction of off-resonance effects and subject movement (Andersson and Sotiropoulos (2016)). The framework is based on aligning and comparing a Gaussian Process based generative model (Andersson and Sotiropoulos (2015)) to the observed data. The extended version, presented here, involves an additional step as part of comparing the generative model to each observed slice, where the slice can be labeled as an outlier based on a summary statistic.

We show that the proposed method has a very high sensitivity and specificity for detecting outliers, and that replacing them by non-parametric predictions almost completely corrects errors in derived parameters.

3 THEORY

3.1 Origin of the dropout

The effects of bulk motion have been described in the case of multi-shot sequences by Anderson and Gore (1994), Trouard et al. (1996), Atkinson et al. (2000) and Norris (2001). For single shot sequences the effects have been described by Wedeen et al. (1994) and there is a very clear description in Storey et al. (2007). In section S.1 in the supplementary material we give an intuitive explanation of how subject movement can cause signal dropout.

In brief:

- Signal dropout is caused by gross movement (subject movement or pulsatile movement of the brain) coinciding temporally with the diffusion encoding.
- Only movement that has a rotational component can cause dropout.
- The severity of the dropout is a function of the magnitude and rotational axis of the gross movement, the imaging plane and the diffusion gradient direction.

- Unlike in the case of distortions the phase encode(PE)-direction has no special relevance, with one caveat: If partial k-space sampling is used the "margin" is smaller in the PE-direction and for a given k-space translation the signal is more likely to end up outside the sampled window. *I.e.* if one for example uses 6/8 partial k-space sampling in one direction the signal only needs to be translated by FOV/4 for it to fall outside of the sampled window compared to FOV/2 for the other directions.

3.2 How to detect the dropout

When visually inspecting a set of diffusion images any sizeable dropout will be noticed as a discrepancy between what one sees and what one "expects" to see. If one for example views a range of slices constituting a volume acquired with a given diffusion gradient it is immediately obvious if one slice has a much lower intensity than the neighboring slices. Hence, one way to detect dropout is to compare an "expectation" of what the signal should be to the observed signal. That "expectation" can be derived from an assumption about neighbouring voxels or from an assumption about neighbouring points in Q-space (diffusion gradient direction and b-value). In the present paper we assume the latter and utilise a Gaussian process (Andersson and Sotiropoulos, 2015) to derive an expectation for what a slice "should" look like.

In brief, a Gaussian process can be used to make predictions about some continuous function $f(\mathbf{x})$ given some "training data" consisting of a set of $(\mathbf{x}, y)$ pairs without having to specify a parametric form for $f(\mathbf{x})$. It does so by assuming the existence of a function $k(\mathbf{x}, \mathbf{x}')$ that gives the covariance of the function values for any pair $\mathbf{x}$ and $\mathbf{x}'$. The form of $k(\mathbf{x}, \mathbf{x}')$ is typically such that the function $f(\mathbf{x})$ varies smoothly as $\mathbf{x}$ is changed. Specifically for diffusion MRI data we recently suggested (Andersson and Sotiropoulos, 2015) a covariance function that combines a "spherical" covariance function (Wackernagel, 2003) for the angular direction and a squared-

exponential (Rasmussen and Williams, 2006) in the b-value direction. It is parameterised by a small number ($3 + N$ where N is the number of shells) of hyperparameters that are learned directly from the data, and can provide predictions for single- or multi-shell data.

Importantly the Gaussian process can give a prediction $\hat{y}$ for any measurement point $\mathbf{x}$ (just as a parametric model like *e.g.* the tensor). The specific form of $\mathbf{x}$ is $[\theta\ \phi\ b]$ and $\hat{y}$ signifies the expected signal in a voxel of the diffusion weighted image acquired after the application of a gradient with direction (θ, ϕ - the spherical coordinates) and b-value b. Formally the prediction $\hat{y}$ for a point $\mathbf{x}^*$ given the set of points $\mathbf{X}$ and observed values $\mathbf{y}$ is

$$\hat{y}(\mathbf{x}^*) = \mathbf{k}(\mathbf{x}^*, \mathbf{X}) \left(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y} \tag{1}$$

where $\mathbf{x}^*$ may or may not be part of $\mathbf{X}$ and where σ is one of the hyperparameters that are learned from the data (see Andersson and Sotiropoulos (2015) for more details).

Hence, a good candidate for detecting an outlier would be $d = y - \hat{y}$ where a negative value would indicate less than expected signal and a potential outlier. If we denote a voxel by its index i we can use $d_i = y_i - \hat{y}_i$ to denote the deviation from the expected signal in voxel i. Since the dropout will pertain to an entire slice or, in the case of pulsatile movement of the brain, to a sizeable subset of the voxels in a slice, it makes sense to use a summary statistic for the entire slice. We will denote this

$$d_{gs} = \frac{1}{n_s} \sum_{i \in s} y_{gi} - \hat{y}_{gi} \tag{2}$$

where s denotes slice s, g denotes volume g, n_s denotes the number of brain-voxels in s and $\hat{y}_{gi}$ denotes the expected signal in voxel i of volume g.

Typical numbers of slices (and hence distinct d_{gs} values) would range from ~a couple of thousand up to more than 50000.

From the collection of d_{gs} we calculated the mean ($\overline{d}$) and an estimate of the voxel-wise

standard deviation from

$$\sigma_d^2 = \frac{1}{N_s - 1} \sum_{s=1}^{N_s} \frac{n_s}{N_g - 1} \sum_{g=1}^{N_g} (d_{gs} - \overline{d})^2 \tag{3}$$

where N_s is the number of slices per volume and N_g is the number of volumes. We can then convert each d_{gs} value to a z-score by

$$z_{gs} = \frac{\sqrt{n_s}\left(d_{gs} - \overline{d}\right)}{\sigma_d} \tag{4}$$

where it should be noted that $\overline{d}$ will be very close to zero. A decision can then be made for any z_{gs} by comparing it to a determined (arbitrary) limit such as for example 3 or 4 standard deviations.

3.3 Dropout and spatial pre-processing

Image registration is needed in dMRI to correct for subject movement and EC-induced distortions. It seems reasonable to assume that if the signal is "missing" in a slice it will have a deleterious effect on image registration and may affect its accuracy. Conversely if two volumes are not properly aligned it is easily realised that a large δ_{gs} is not necessarily an indication of signal dropout, but could be due to the lack of registration. Hence, one cannot correct for movement and distortion in the presence of dropout nor can one detect outliers if there are uncorrected distortions and subject movement, which suggests that they all need to be corrected simultaneously.

A simultaneous correction is complicated by the way most EC correction is performed: by aligning all images to some reference (typically an image without diffusion weighting that is unaffected by EC) as shown in figure 1. It means that when an image affected by a slice dropout is transformed into reference space where a prediction could be made, that slice might potentially be slanted in both an xz- and and yz-view. This would make it more difficult to identify it as

an outlier since one can no longer use a statistic based on the entire slice and there will be interpolated voxels which are only partially affected by the dropout.

We recently suggested a method for estimation of and correction of EC-induced distortions and subject movement where a Gaussian Process model in reference space is transformed into the (unique) space of each of the observed volumes (Andersson and Sotiropoulos, 2016). The comparison that drives the registration is then performed in the scan space. As shown in figure 2 this formulation lends itself very easily to outlier detection since the observed scan is compared directly to the model prediction without any resampling.

3.4 The algorithm

The algorithm for estimation and correction of eddy currents and movement is described in detail in Andersson and Sotiropoulos (2016) and here we focus on the modifications that allow us to also detect and replace outliers. A schematic outline of the algorithm is presented in figure 3 divided into two steps, i) The prediction step and ii) The estimation step. One iteration of the method consists of performing both steps.

The difference between the original algorithm and the modification we propose in the present paper is highlighted in red in figure 3 and consists of keeping a list of outliers that is updated as part of the estimation step. The outlier list keeping can be summarized as shown in Algorithm 1

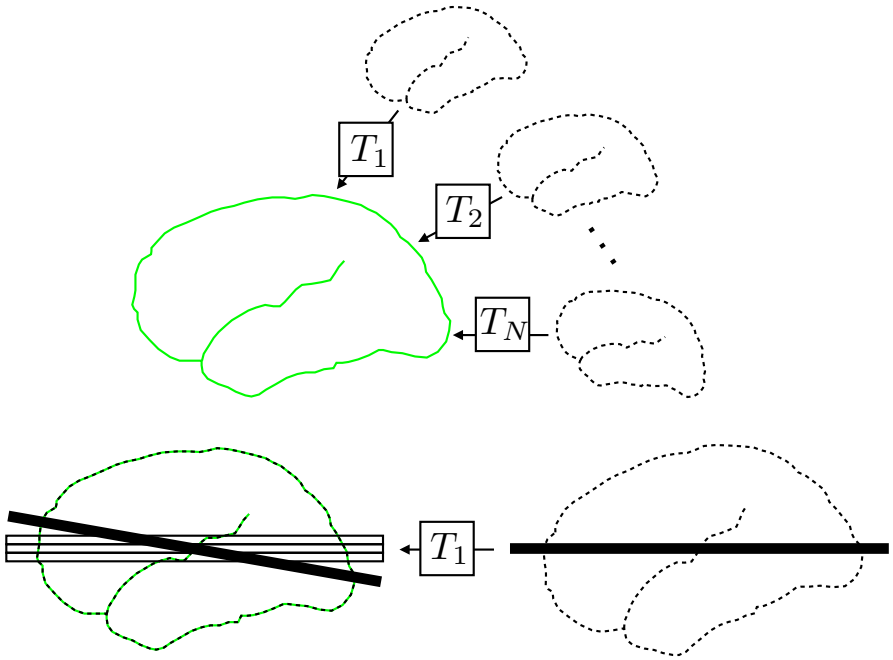

Figure 1: The top row shows the principle behind many image registration based EC/movement correction methods. A reference (which might be a specific scan or some function of many scans in a "reference space", shown with a gren outline) is used as a target to which the individual scans (shown with dashed outlines) are registered through one transform T_i per scan. The lower row demonstrates the case when there is an outlier slice (solid, thick line) in one of the scans. As it is registered to the target the outlier slice no longer lines up with a slice in the reference, which makes it harder to detect.

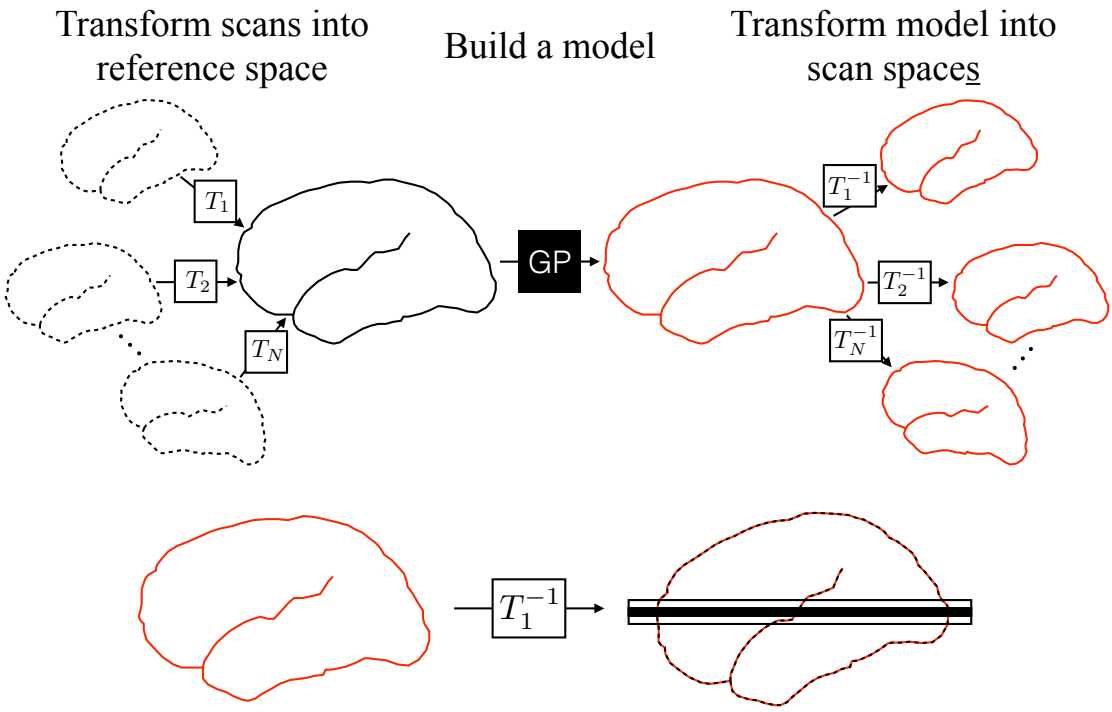

Figure 2: The top row shows how each scan is transformed (with its own unique transform) into a single reference space and how a Gaussian Process model is built in that space. For a given volume/gradient a prediction (shown in red) is made in reference space and transformed into the pertinent scan space using the inverse of the transform for that volume (bottom row). The prediction is now in scan space and the comparison between prediction and observation can be performed without any resampling of the actual scan data.

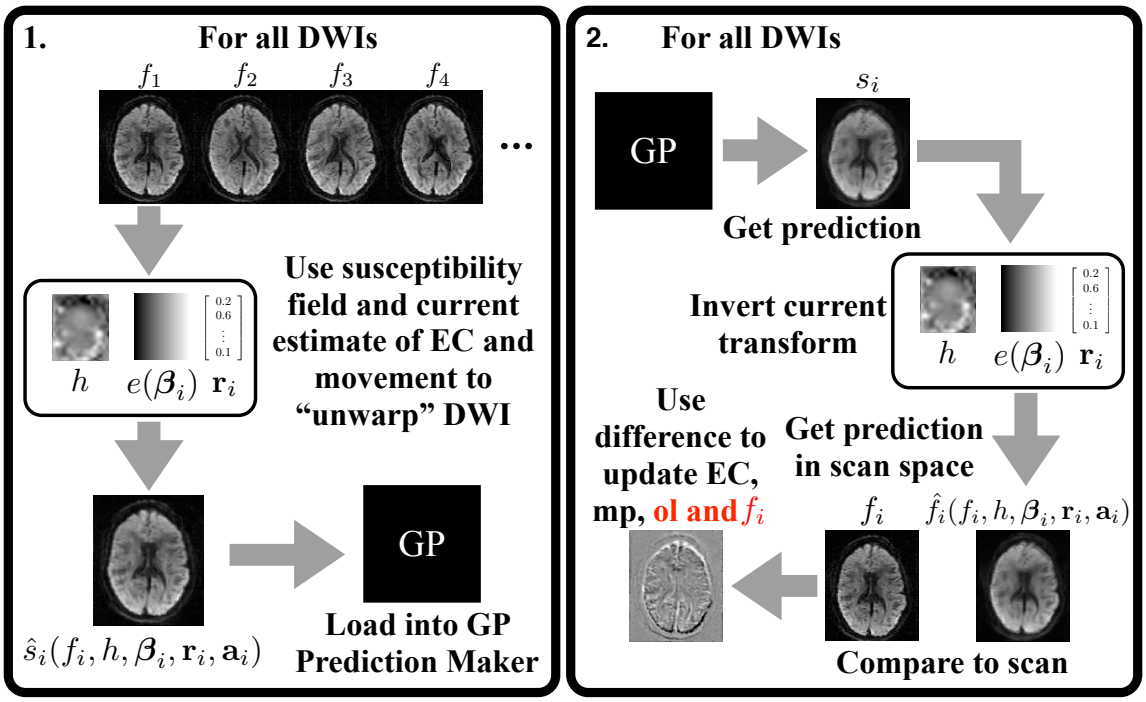

Figure 3: This figure attempts to graphically demonstrate the joint distortion, movement and outlier algorithm. The difference to that presented in Andersson and Sotiropoulos (2016) is highlighted in red and consists of an additional step at the stage of comparing prediction and observation. If the observation slice is deemed to be an outlier it is put on a list of outliers (ol) and is replaced by the prediction. The details of the replacement are given in algorithm 1.

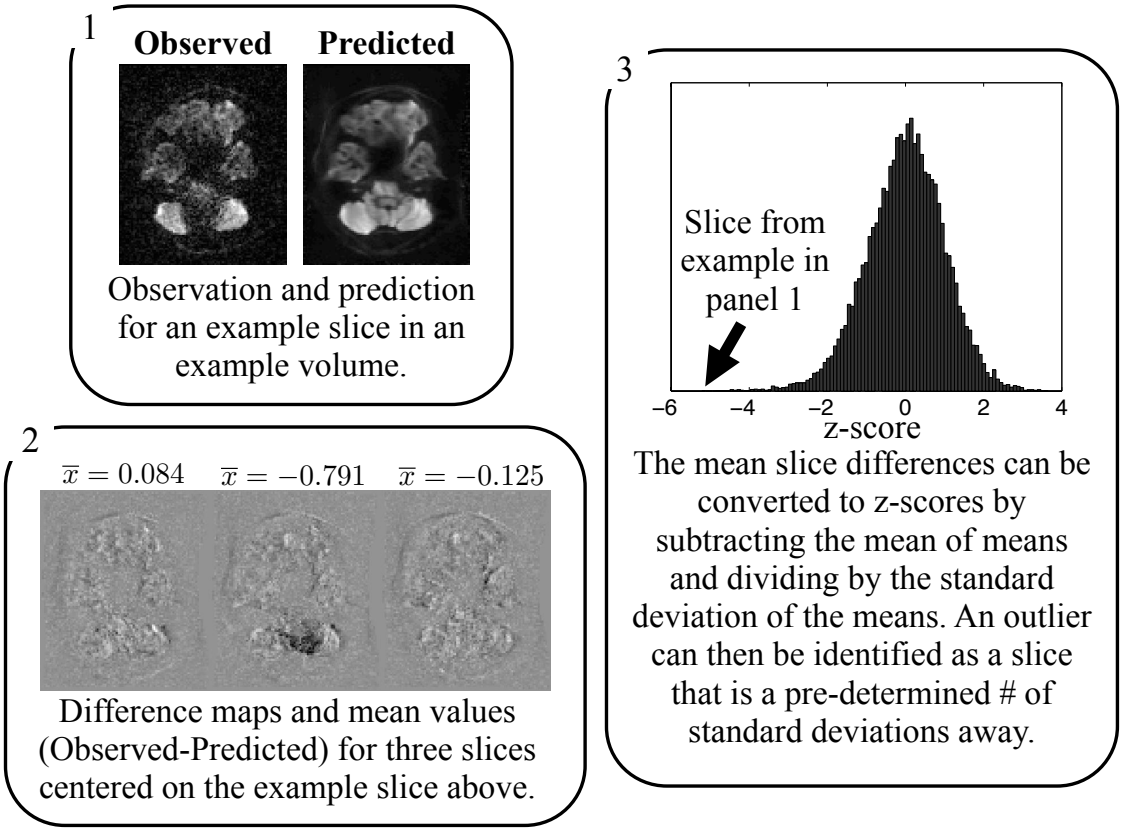

Figure 4: This figure is based on data from an HCP subject and shows a schematic of how an outlier is detected. For each slice in each volume a prediction is made and compared to what was observed (panel 1). For each pair of observation–prediction the voxelwise mean of differences is calculated (panel 2). From the set of all such differences the mean and standard deviation is calculated. Then the number of standard deviations between the slice mean and the average slice mean (z-score) is calculated. The decision to reject a slice is binary and is based on the z-score exceeding some pre-defined threshold.

```
Initialize:
Make an empty list of outlier slices
Put all original slices into list of used slices
for Every eddy iteration do
    for Every slice in every volume do
        Make prediction based on used slices
        Calculate difference between prediction and original slice
        if Difference indicate an outlier then
            if Slice is already in outlier list then
                Replace used slice by prediction
            else
                Make new prediction without current volume
                Replace used slice by new prediction
                Add original slice to outlier list
            end
        else
            if slice is on outlier list then
                Put original slice back into used slices
                Remove slice from outlier list
            end
        end
        Calculate difference between prediction and used slice
        Use difference to update movement and EC parameters
    end
end
```

Algorithm 1: This algorithm details the outlier part of the joint distortion, movement and outlier correction algorithm.

Importantly, as can be seen from the pseudo-code above, a slice can be labeled an outlier in one of the first iterations and then be rehabilitated and brought back in a later iteration. This is important since a slice may appear as an outlier due to uncorrected movement or EC-induced distortions, but may not once the movements and distortions have been corrected. Conversely, in an early iteration, sensitivity to detect outliers may be poor as the distribution of slice differences (c.f. right panel of figure 4) is quite wide because of movements and distortions. But as these are corrected the distribution becomes more narrow, the sensitivity improves and new slices will be deemed as outliers.

It is our experience that this algorithm is very robust and we have not seen it fail to converge on any of the simulations or experimental data presented in the paper.

3.4.1 Outlier detection for simultaneous multi-slice acquisitions

With modern multiband (MB) acquisitions, when several slices are acquired in groups following a single diffusion encoding (Moeller et al. (2010), Setsompop et al. (2012)) any subject movement will affect the signal in all of those slices. We therefore suggest the group (of slices) as the unit for detecting an outlier. It is trivial to adapt equations 2 to 4 by simply replacing each s by an m (for $m$ulti-band group). This should further increase our ability to detect an outlier (group), *i.e.* make the method more sensitive.

However, when the true cause of the dropout is pulsatile physiological movement (leading to a local "rotation"), rather than subject movement, it can cause signal loss only/predominantly for one slice even for multi-band acquisition. To resolve this we have implemented a heuristic that combines the slice- and group-wise tests. We refer to this last method as "combined" (combining group- and slice-wise detection).

3.5 What to do with the dropout

When a dropout has been detected the next question is what to do about it. One option would be to discard the detected slice, but there are some practical problems associated with that. The end result of the corrections is a dataset in a single (reference) space which means that all of the individual diffusion weighted images (dwis) will have been resampled (including out-of-plane rotations) into that space. Hence a "discarded" slice will now be a discarded oblique plane with ambiguous edges, and for which one would have needed the information from the discarded slice for the interpolation.

Instead, we propose to replace the original slice with the prediction made by the Gaussian Process (GP) and then proceed as if there was no missing slice. This is equivalent to replacing a data-point with its expectation (given the GP), which is the value that would not affect the

predictions/inference of the GP at all. That means that it is independent of whatever biophysical model will subsequently be used to analyse the data and will therefore not bias results towards any one specific model. The quality of that prediction, in terms of similarity to some unknown "truth", will depend on how densely the diffusion sphere has been sampled. Unlike the distortion correction estimated by `eddy` it does not matter if diffusion has been sampled on the whole or the half sphere. More details about the Gaussian Process generative model can be found in Andersson and Sotiropoulos (2015).

4 MATERIAL AND METHODS

4.1 Implementation

The method described in the present paper has been implemented in C++ as part of the `eddy` tool (`http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/EDDY`) in FSL (see Smith et al. (2004) and `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`). When we use the term `eddy` we refer to that implementation of joint movement, distortion and outlier correction.

4.2 Simulations

4.2.1 Data

We used the POSSUM MRI-simulator (Drobnjak et al. (2006), Drobnjak et al. (2010)), extended to simulate diffusion weighted images (Graham et al. (2016)). POSSUM is a highly realistic MRI simulator that simulates data in k-space by solving Bloch's and Maxwell's equations. This ensures that images and their artefacts capture the key features of their real-world counterparts. The object (brain) was based on a tissue type segmentation of a subject from the HCP cohort (Van Essen et al. (2013)) followed by tissue type based assignment of T1, T2 and T2*. The

diffusion properties was based on a spherical harmonics fit to the subject's diffusion data.

Data was simulated with known eddy currents, subject movement and signal dropout. Simulations were divided into slice-wise and group-wise outliers, where a group corresponds to a set of slices acquired for the same excitation and diffusion weighting as part of a multi-band acquisition. For all simulations there was "ground truth" dataset that contained no noise, movement or outliers. A simulated dataset consisted of 12, interspersed, b=0 volumes, 32 $b = 700\text{s/mm}^2$ volumes and 64 $b = 2000\text{s/mm}^2$ volumes. The matrix size of each volume was $72 \times 86 \times 55$ with a $2.5 \times 2.5 \times 2.5$mm voxel size. The directions for each non-zero shell were optimised on the half-sphere using Coulomb forces (Jones et al. (1999)), and then randomly sign-swapped to achieve a "reasonable" distribution over the whole sphere. Two different levels of subject movement and two SNR levels were used. The movements were "realistic" in that they were taken from the estimated motion parameters of actual studies, one set for a "good" subject and one for a "bad" subject. The two SNR levels were defined for the b=0 volumes and were 40 and 20, corresponding to "normal" and "poor" SNR. Outliers were generated for random slices in random volumes with a multiplicative "dropout factor" randomly drawn between 0.1 and 0.9, where zero corresponds to complete dropout and 1.0 to no dropout. The total number of outliers was chosen to be commensurate with what is "typically" seen in datasets with "slightly difficult" subjects such as in the Whitehall Imaging substudy ((Filippini et al., 2014)) or in the study by Krogsrud et al. (2016). This corresponded to 155 ± 25 outliers per dataset for these simulations with 96, diffusion weighted volumes of 55 slices each. Not all outliers occurred in slices with a sufficient number of brain-voxels for estimation of outlier slices, so within the slices actually used there were 126 ± 21 outliers. Data was also simulated with the same EC-distortions, motion parameters and noise levels, but without any outliers. This was to enable the isolation of the effects of the outliers from the other factors that degrade the data, and also to be able to assess

any negative effects from false positives (the erroneous labelling of a slice or voxel as an outlier) when applying outlier detection to data with no outliers.

4.2.1.1 Slice-wise outliers

This corresponds to the usual case where each slice is excited and diffusion weighted separately, and hence where a movement during the diffusion weighting affects a single slice. The outliers slices were drawn individually from a random diffusion weighted volume followed by a random slice within the volume.

4.2.1.2 Group-wise outliers

In multi-band (MB) imaging several slices are excited and diffusion weighted simultaneously and any movement affects them all. Hence, potentially it is better to assess outliers as "outlier groups" than outlier slices. For this case outliers were created by drawing a random diffusion weighted volume followed by a random group (of slices) within the volume where all slices within the group were multiplied by the same factor. Our framework has the option to use either slice-wise or MB-group-wise statistics and outlier detection, or a combination of both.

4.2.1.3 Increasing frequency of outliers

The data above were simulated with an outlier frequency of 3%, which corresponds to what is typically seen in data. These 3% were realised by a 30% chance of an outlier in each volume and a 10% chance of any slice being an outlier in a volume with outliers. For this simulation both those probabilities were increased by $\sqrt{n}$ (where n was $1, 2, \ldots, 6$) such that the total frequencies of outliers were 3, 6, 9, 12, 15 and 18%. This simulation was performed only for the level of movement corresponding to a "good" subject and for slice-wise outliers.

4.2.2 Analysis

4.2.2.1 The effect of outliers on estimated movement and EC parameters

As outlined in section 3.3 it seems likely that the presence of outliers in the data will adversely affect our ability to accurately estimate subject movement and EC-induced distortions. For the simulated data the "true" geometric distortions are available as displacement fields that summarise the effects of movement and EC distortions so these can be compared to the corresponding fields estimated by `eddy` to assess the accuracy of the latter (Graham et al. (2016)). For this analysis we used the sum-of-squared differences between the true and estimated displacement fields (summed over all brain-voxels) as a summary statistic of the registration error. We used `eddy` on the simulated data with slice-wise outliers (section 4.2.1.1) using outlier replacement at threshold levels of 3, 3.5, 4 and 4.5 σ and also without any outlier replacement.

4.2.2.2 Comparison to RESTORE

RESTORE is a previously published method (Chang et al. (2005), Chang et al. (2012)) that fits a tensor model non-linearly to each voxel and detects outliers for that voxel based on the residuals (one for each volume) of that fit. There are a number of other methods available that we could have compared against (for example PATCH, Zwiers (2010) or DTIPrep, Oguz et al. (2014)). The choice of RESTORE was based on it being amongst the first published methods and being part of two popular packages for analysis of diffusion data (CAMINO, Cook et al. (2006) and TORTOISE, Pierpaoli et al. (2010)). For this comparison, all processing following `eddy` was performed using the CAMINO software package (Cook et al. (2006)) (but also see Supplementary Material for comparisons using RESTORE as implemented in TORTOISE, Pierpaoli et al. (2010)). All simulations were pre-processed in two ways:

`eddy` EC-induced distortions and subject movements were corrected using `eddy`.

`eddy` with OLR EC-induced distortions, subject movements and outliers were corrected using `eddy` with simultaneous outlier rejection.

For this comparison the outlier detection in `eddy` considered both positive (increased signal) and negative outliers in order to be comparable to that of RESTORE since the implementation in CAMINO does not support the "informed" detection described in Chang et al. (2012). In order to isolate the effects of the outlier detection, `eddy` used outlier replacement internally to ensure that the estimation of movement and EC-induced distortions did not differ between options `eddy` and `eddy` with OLR. *I.e.* it used the algorithm shown in algorithm 1 for the estimation of the movement and distortion parameters, but for the option `eddy` those parameters were applied to the original data with the outliers left in place. A single diffusion tensor model was fitted non-linearly (Jones and Basser (2004)) as implemented in the CAMINO software package (Cook et al. (2006)) to data pre-processed in all three ways, as well as to the noise free "ground truth" data. In addition, for data pre-processed according to `eddy`, single tensors were fitted using RESTORE (Chang et al. (2005)) as implemented in CAMINO (Cook et al. (2006)). This processing stream will be referred to as `eddy`+RESTORE. For the RESTORE analysis we estimated the noise using the "for multiple b=0 images" option of the `estimatesnr` command of CAMINO. These estimates were in close agreement with the SNR used to simulate the data (within ± 1 of the tentative values (20 or 40)). We calculated the correlation (Pearson's r) between "ground truth" derived FA and MD and that derived from the processed data. This was performed both for all brain-voxels (denoted r in figure 7) and for all white matter voxels only (denoted rw in figure 7).

4.2.2.3 False positives and negatives

A false positive is when a slice or group is labelled as an outlier when in fact it is not. A false negative is an outlier slice that is not detected. For this analysis outlier detection (using `eddy` with OLR) was performed for four different thresholds for defining an outlier: 3, 3.5, 4 and 4.5 standard deviations away from the mean slice-difference. Only "negative" outliers (*i.e.* outliers associated with signal loss) were considered. Outlier detection was only attempted for slices with 250 or more brain-voxels (as defined by a user supplied mask). Outliers in other slices were ignored, and hence did not count towards false negatives. For each movement type ("good" or "bad" as defined in section 4.2.1 above) and SNR level (20 and 40) ten different realisations of random outlier slices were used.

For the analysis of false positives we used the simulated data without outliers and analysed it using slice-wise, group-wise and combined outlier detection. The false negatives were assessed using data with slice-wise and data with group-wise simulated outliers. Data with slice-wise outliers were analysed with slice-wise and combined outlier detection, and data with group-wise outliers were analysed with group-wise and combined outlier detection.

The data without outliers were also used to assess false positives with RESTORE. Because of the difficulty of clearly defining an outlier after movement correction and interpolation no attempt was made to estimate the false negative rate for RESTORE.

4.2.2.4 The effect on derived diffusion parameters

Different methods for estimating the diffusion tensor are differently sensitive to the presence of outliers caused by signal loss. This is due to the different weighting of the error that is implicit in the different methods (see *e.g.* Chang et al. (2005)). For this analysis tensors were fitted to the data using the CAMINO implementations of

OLS Ordinary linear least-squares on log-transformed data

WLS Weighted linear least-squares on log-transformed data (Jones and Basser (2004)).

NLS Non-linear least-squares (Jones and Basser (2004)).

This was done for the pre-processing streams `eddy` and `eddy` with OLR in section 4.2.2.2 above, but for this analysis only detecting and rejecting "negative" outliers. For each movement type ("good" or "bad" as defined in section 4.2.1 above) and SNR level (20 and 40) ten different realisations of random outlier slices were used. The outcome was the correlation (Pearson's r) between "ground truth" derived FA and MD and that derived from the processed data.

4.2.2.5 The effect of frequency of outliers

A similar analysis to in the previous paragraph was performed, but for this analysis only with a threshold of 4σ. Data with SNR levels (20 and 40), movement corresponding to a "good" subject, outlier frequencies of 0, 3, 6, 9, 12, 15 and 18% was used and for each case ten different realisations of noise and random outlier slices were used. The outcome was the correlation (Pearson's r) between "ground truth" derived FA and that derived from the processed data.

4.3 Whitehall Imaging data

4.3.1 Data

These data are from community-dwelling older adults and have been described in great detail in Filippini et al. (2014). Images were acquired on a 3T Siemens Magnetom Verio system with a 32-channel receive head coil. Diffusion encoding was performed using monopolar Stejskal-Tanner gradients. A single shell with 60 unique directions and a b-value of 1500 was acquired along with 5 interspersed $b = 0$ volumes. The resolution was 2mm isotropic, the imaging matrix

was 96×96×64 voxels. Phase-encoding was in the anterior→posterior direction and a factor of 2 in-plane acceleration (GRAPPA) was used. An additional single $b = 0$ volume with phase-encoding posterior→anterior was acquired to allow for estimation of the susceptibility-induced off-resonance field.

Data was used from eight subjects who were identified as "having moved a lot".

4.3.2 Analysis

For these datasets there was no "ground truth", hence the analysis consisted of manual identification of outlier slices. First, the sagittal and coronal views of each volume were inspected for signs of "stripy appearance" in the z-direction. Each volume with signs of "stripy appearance" was stepped through in the z-direction whilst viewing the transversal slices. If a slice appeared "surprisingly dark" it was further inspected to determine if it was an outlier. This process was performed twice, once stepping in bottom to top and once in top to bottom direction.

Subsequently, `eddy` was run with and without outlier replacement and the data was revisited with the automated outlier report in hand to verify the outliers.

5 RESULTS

5.1 Simulations

In figure 5 we show an example of the simulated data. It demonstrates the problem alluded to in figure 1, where after reorientation an outlier slice has turned into oblique bands where most voxels are a mixture of lost and preserved signal.

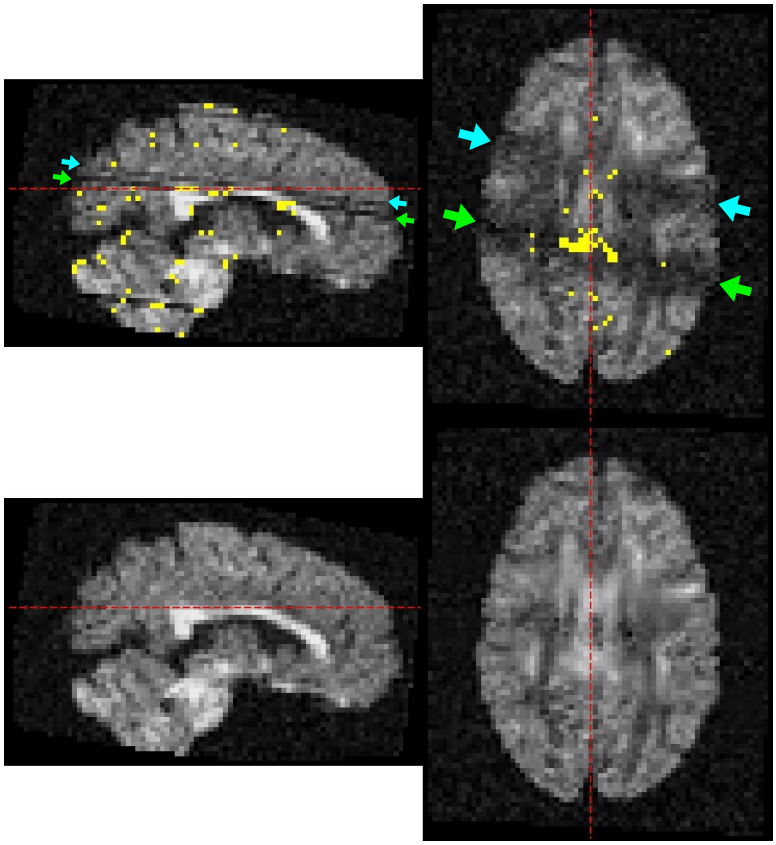

Figure 5: An example of a b=2000 volume from the simulations. This is a volume late in the series where the movement has accumulated to rotations of 7.3 and 1.5 degrees around the x- and y-axes respectively. The top row shows the results without outlier replacement and the voxels labeled as outliers by RESTORE (in yellow). Note in the sagittal view how the outlier slices are now thin oblique bands (indicated by arrows) through the volume and that they manifest as wide oblique bands (indicated by arrows with corresponding color) in a transversal view. All voxels in one of those bands are, through the interpolation, a mixture of outlier and non-outlier voxels. This makes it more difficult to detect (or even define) an outlier (*c.f.* figure 1). The bottom row shows the same slices from the same data when outlier detection and replacement was performed.

5.1.0.1 The effect on estimated movement and EC parameters

Figure 6 shows a typical example of registration error for each volume in one realisation of the simulated data. It can be seen that for most of the volumes there is no appreciable difference when performing outlier rejection compared to when not. The volumes with a clear difference are all characterised by a larger than average number of outlier slices (the two volumes with ten and nine outliers had the largest numbers). In those cases, outlier detection has helped improve the estimates of the distortions and subject movements. On the other hand there are several volumes (not indicated in figure 6) that do not exhibit an increased registration error despite having four or more outlier slices. Our interpretation is that it is not only the number of outliers that determine the effect on registration accuracy, but also the location of these slices in the volume.

The average registration error was consistently higher when not performing outlier rejection, but because it mainly affected a handful of volumes the effect on the mean was very small (table 1).

5.1.0.2 Comparison to RESTORE

Figure 7 shows an example of correlations against "ground truth" FA for the small movement case with SNR 40. When looking at the scatter plots for the `eddy`+RESTORE case it can also be seen that it has many voxels in the low FA range where the FA has been overestimated, which leads to worsened correlation with "truth" when estimated across all brain-voxels. RESTORE performs better when limiting the calculation of the correlation with "truth" to white matter voxels (rw in figure 7), but is still inferior to the outlier replacement within `eddy`.

Figure 8 shows results pooled across all ten realisations for the two levels of movement and SNR.

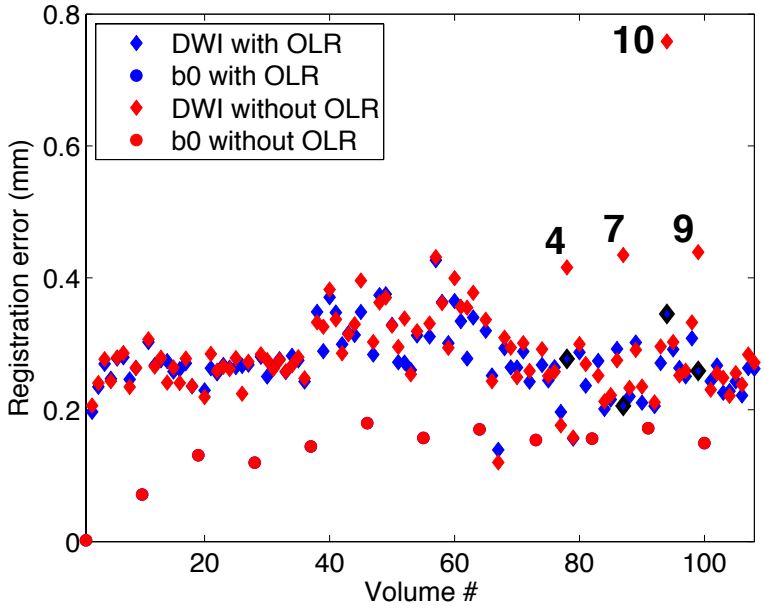

Figure 6: Registration error was defined as the average voxelwise misplacement averaged across all brain-voxels. Registration with outlier rejection is shown in blue and without in red. The numbers next to four of the red markers are the # of outliers in those volumes (the average number of outliers per dwi volume is 1.7). The corresponding points *with* outlier rejection have been framed in black to make them easier to spot. There were other volumes with 4 or more outlier slices (not indicated) that did not show an increased error. The simulated data used for this figure had an SNR of 20 and subject movement corresponding to a "good" subject's (see section 4.2.1). A threshold of 4σ was used to define an outlier slice. There are no blue dots visible in the figure. That is because there are no outliers among the b=0 volumes so they coincide exactly with the red dots.

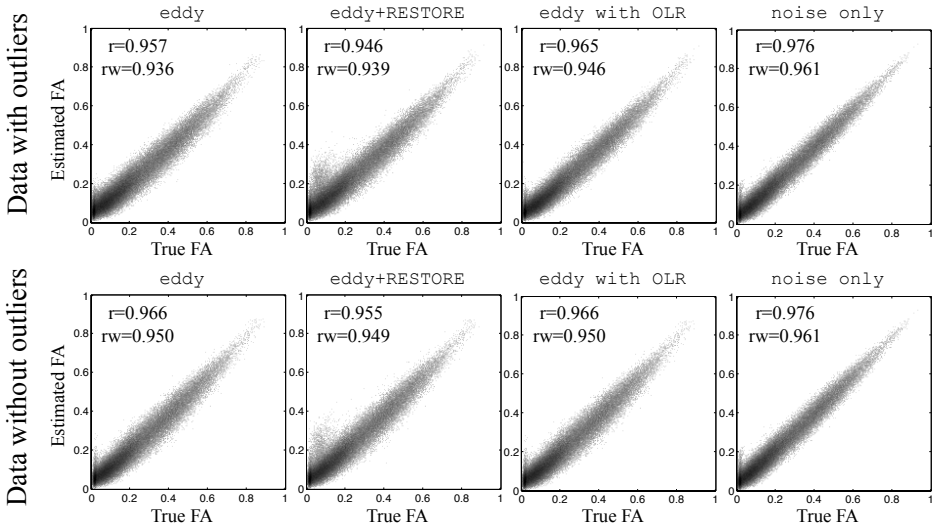

Figure 7: This figure shows the voxelwise correlation between estimated FA and "ground truth" for the simulated data. This example represent one realisation of the "small movement" case with SNR 20. The top row shows the results when data contained outlier slices and the bottom row when it did not. Panels from left to right are i) `eddy` with no outlier correction, ii) `eddy` followed by RESTORE outlier correction, iii) `eddy` with outlier correction (4σ) and iv) data with noise only (*i.e.* the "ground-truth" correlation in the presence of noise). The "scatter plots" have been created by binning all pairs of voxels in a 200×200 matrix and displaying the log of the resulting images as a gray scale. The r-values are the voxel wise correlations calculated over all brain-voxels and the rw calculated for white matter voxels only.

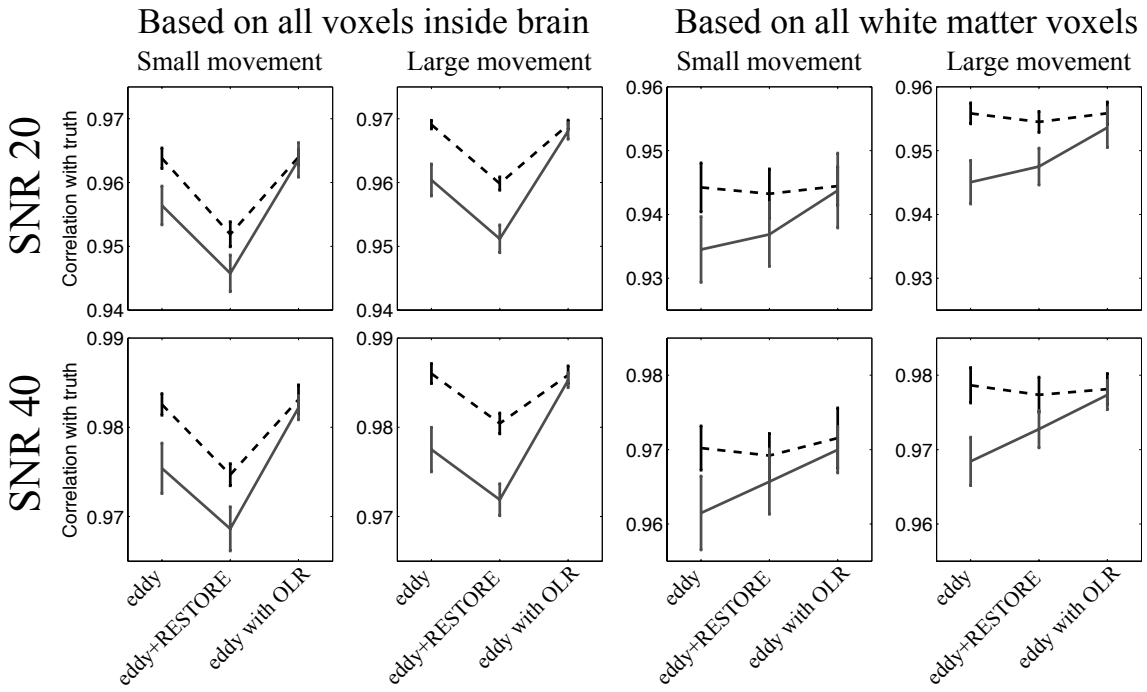

Figure 8: This figure shows the voxelwise correlation between estimated FA and "ground truth" for the simulated data. Each point represents the correlation across all brain-voxels (left two columns) or across all white matter voxels (right two columns) averaged over ten realisations. The error bars represent ± one standard deviation across those ten realisations. The dashed black lines represent the cases where there were no outliers in the data, and the solid grey lines when there were outliers.

Movement:	Small		Large	
SNR:	20	40	20	40
Without OLR:	$.279 \pm .015$	$.242 \pm .014$	$.269 \pm .004$	$.248 \pm .005$
With OLR:	$.264 \pm .013$	$.239 \pm .009$	$.259 \pm .004$	$.239 \pm .003$
Difference	.015 (2.42*)	.003 (0.58)	.010 (5.42*)	.008 (4.63*)

Table 1: Error and difference in registration error when performing compared to not performing outlier rejection as part of the registration process. The numbers represent average (over all dwi volumes and over the ten realisations) error in mm. The first two rows show the error without and with outlier replacement and the third row the increase in registration error. The numbers in parentheses are t-statistics from a two-sample test with unequal variances and the * indicates significance at the 0.05 level.

Figure 8 shows that in the absence of outliers RESTORE consistently decreases the correlation to ground truth, though less so when only considering white matter voxels. When data contained outliers RESTORE still decreased the correlation when considered across all brain-voxels, but improved the correlation for white matter voxels.

In contrast the outlier replacement within `eddy` did not affect results when there were no outliers. This can be seen by comparing the first and the third point of the dashed lines in figure 8. When data contained outliers it was able to correct for those to a level that was comparable to the case when there were no outliers in the first place, as can be seen by comparing the first points of the dashed lines to the third points of the solid lines.

5.1.0.3 False positives and negatives

The false positive rate was assessed from the data without outliers, which means that any slice labeled as an outlier was a false positive. The rate was assessed for four different outlier

Movement:	Small		Large	
SNR:	20	40	20	40
Slice-wise:	.00042	.00016	.00056	.00127
Group-wise:	0	0	0	0
Combined	.00044	.00019	.00060	.00104

Table 2: Observed false positive rate for `eddy` with OLR averaged across ten simulated datasets. An outlier threshold of 3 was used, which means that the tentative false positive rate was 0.00135. When using a higher threshold (3.5, 4 or 4.5) no false positives were observed.

Movement:	Small		Large	
SNR:	20	40	20	40
False positive rate:	$.01125 \pm .00024$	$.02020 \pm .00043$	$.00858 \pm .00013$	$.01349 \pm .00031$

Table 3: This table shows the false positive rate for RESTORE (in the `eddy`+RESTORE chain) for each simulated dataset.

thresholds, 3, 3.5, 4 and 4.5 in datasets with a total of 4320 slices with sufficient brain-voxels and diffusion weighted images. The rate of false positives was lower than the tentative (*i.e.* more conservative) rate for all cases, and only for the lowest threshold (3) did we find any false positives at all. The results for an outlier threshold of 3 are summarised in table 2. The reason that the combined error rate is not always greater than the slice-wise is that there is a stochastic element in the estimation of the hyperparameters on which the predictions depend.

The number of false positive voxels from RESTORE are reported in table 3. For all cases the rate of false positives was considerably higher than the approximate tentative false positive rate of 0.00135 (based on 3σ). Looking at spatial maps (data not shown) of these false positives they are predominantly along edges between cerebrospinal fluid (CSF) and brain.

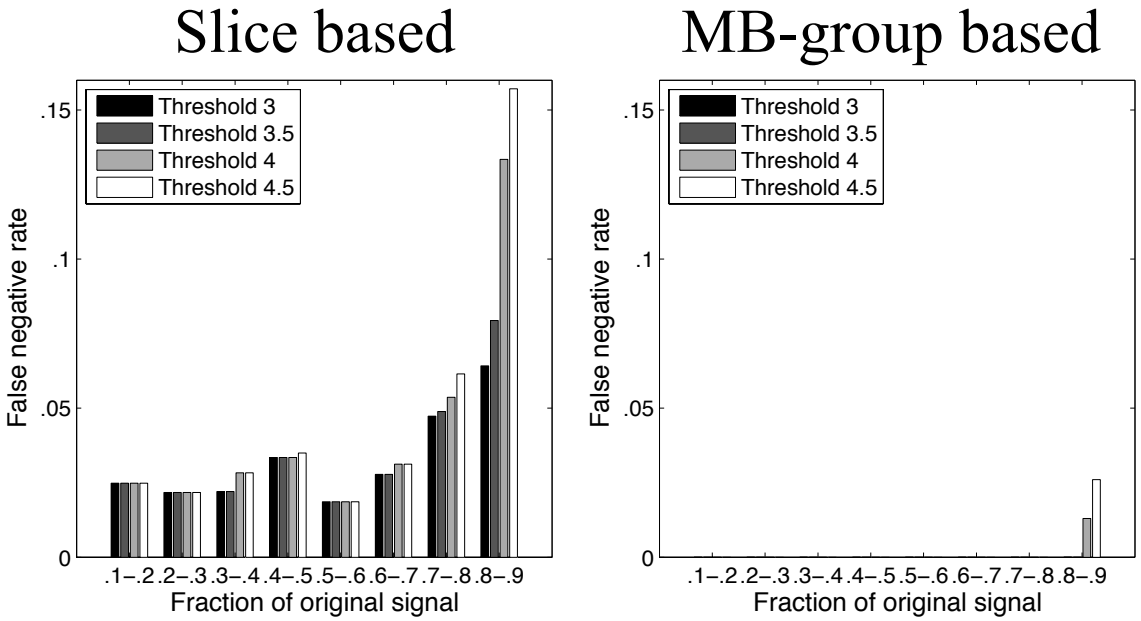

Figure 9: This figure shows the false negative rate vs fraction of original signal intensity (0 is complete signal dropout, 1 is no dropout). In the left panel it can be seen that for fractional signal of 0.1–0.7 the FN rate is relatively constant, and that for fractional signal $> 0.7\times$original signal the FN rate starts to increase. For the MB-group based outlier detection there are no false negatives for fractional signal $< 0.8\times$original signal.

As expected the false negative (FN) rate depended on the threshold used for the detection with a rate of 0.030 for a threshold of 3 and a rate of 0.045 for a threshold of 4.5 (the similarity between the numbers for the rate and threshold is a coincidence). The effect of SNR is surprisingly small with FN rates of 0.0395 and 0.0381 for SNR 20 and 40 respectively when averaged across thresholds. As can be seen in figure 9, it also depends on the "strength" of the outliers such that an outlier slice with 20% of the original intensity is more likely to get detected than one with 80%.

There is also a large effect of the number of brain-voxels in a slice on the ability to detect

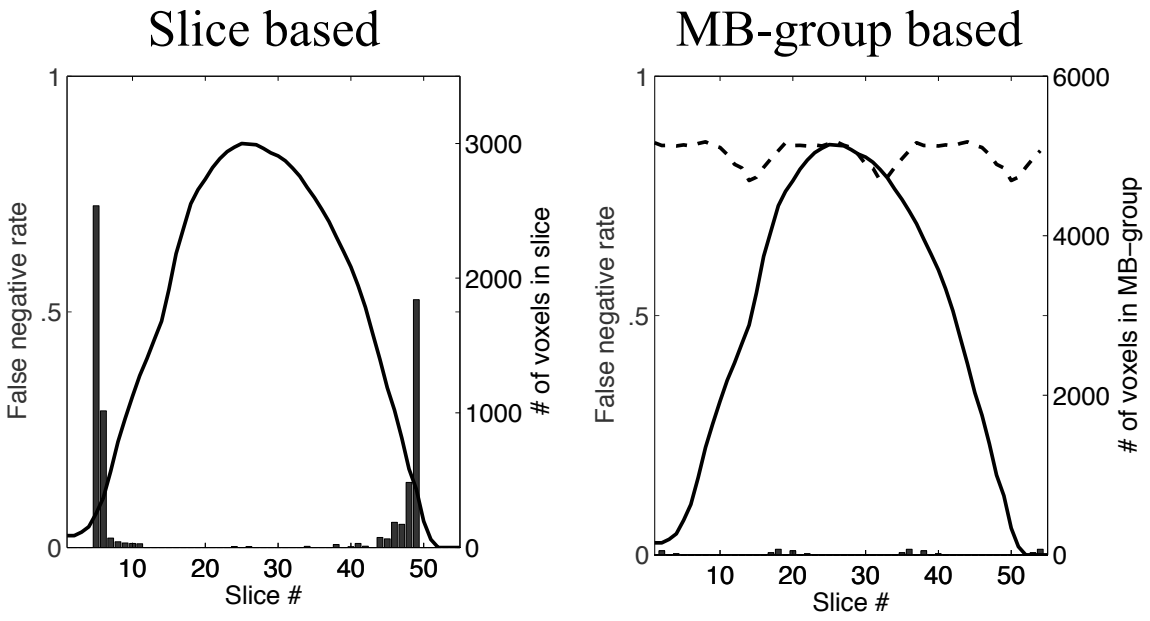

Figure 10: This figure shows the false negative rate (based on 4σ) vs slice number. The bars show the false negative rate per slice (left column) or per MB-group (right column). Also shown is the # of brain-voxels per slice (solid line and right y-axis on left panel) and # of brain-voxels per MB-group (dashed line and right y-axis on right panel).

an outlier in that slice. Almost all the false negatives were in the ultimate slices at the top and bottom (see figure 10). For the top and bottom three slices the average number of voxels were 502, and the average FN rate was 0.29. For the middle 39 slices, the average number of voxels were 2224 and the average FN rate was 0.0040.

5.1.0.4 The effect on derived diffusion parameters

The results from examining the effect of the outlier detection and replacement on the estimated diffusion parameters are presented in tables S3, S4, S5 and S6. A subset of the results are also summarised in figure 11, though the other results in tables are in good agreement with

that subset. Figure 11 demonstrates that there are no deleterious effects from applying the outlier detection to data without outliers, as evident from the horizontal dashed lines for all plots. It also demonstrates that not correcting for outliers (final point of each line) has a negative impact on the estimates and that this impact appears to be greater for weighted linear tensor estimation (WLS, black solid lines) compared to ordinary linear estimation (OLS, grey solid lines). The deleterious effect of the outliers is almost completely reversed by the outlier detection/replacement (as indicated by the proximity of the solid and dashed lines in each plot) and this seems to be independent of the threshold that was used for outlier detection.

5.1.0.5 The effect of frequency of outliers

The effects of increasing frequency of outliers are presented in tables S3 and S5 and in figure 12. It demonstrates that when outlier detection and replacement is performed there appears to be little damaging effect up to an outlier frequency of 10%, after which results start to get worse. In contrast, when the outliers are ignored, the results decline approximately linearly over the examined range. Already at an outlier frequency of 3% are the results as bad, or worse, as when performing outlier detection and replacement on data with 15% outliers.

5.2 Whitehall imaging data

Manual inspection of the data yielded a number of outlier slices for each subject as presented in table 4 together with the corresponding numbers for `eddy`. As can be seen from the table the automatic outlier detection in `eddy` detects considerably more outliers than the manual observer, even when using a threshold of 4. On the other hand there is a strong correlation between the measures (0.89 between Manual and OLR 3 and 0.95 between Manual and OLR 4) which indicated that the difference can just be a matter of sensitivity. This is further supported by a detailed analysis of the results from the "worst" subject: WH_025 (as defined by the

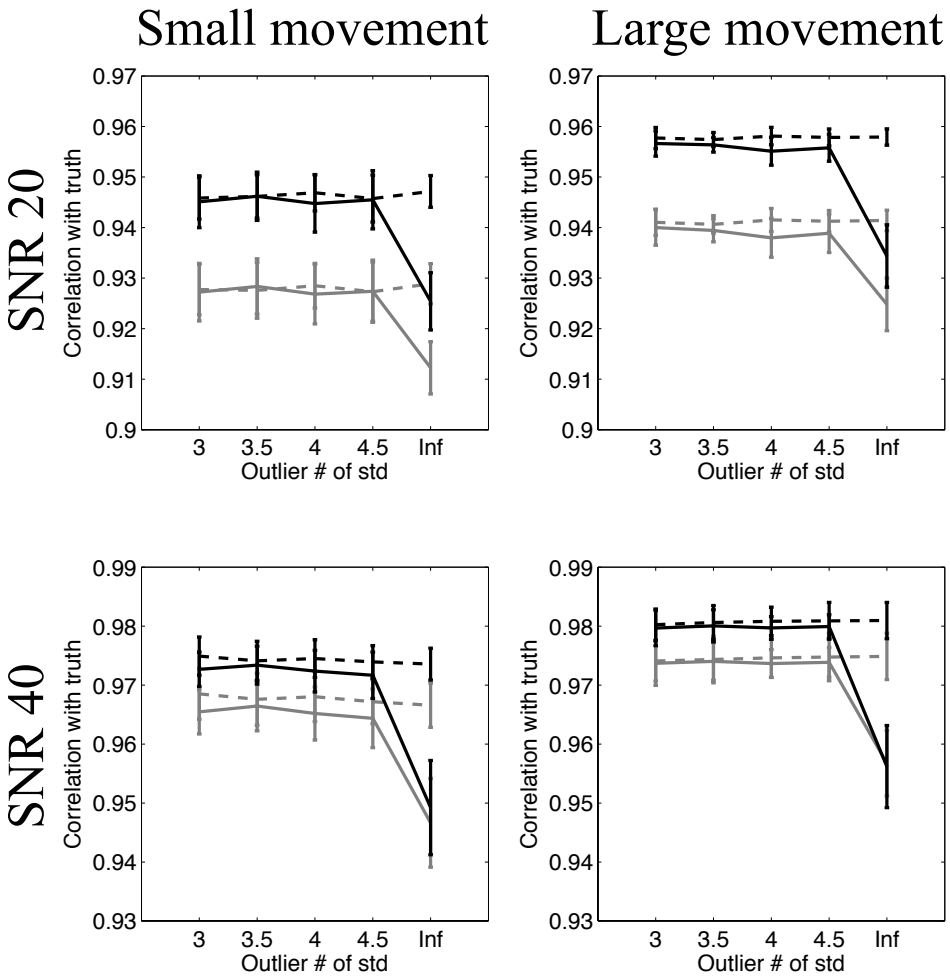

Figure 11: This figure shows the correlation between estimated and "true" FA. The left column is for data simulated with small movement and the right column for large movement. The top row is from data simulated with an SNR of 20 and the bottom row for an SNR of 40. Solid lines are for data with outliers and dashed lines for data without. Black lines represent tensor estimation using weighted linear fit (WLS) and grey lines represent tensor estimation with linear fit (OLS). The "Inf" outlier threshold represents "no outlier detection". The error bars represent one standard deviation across the ten realisations.

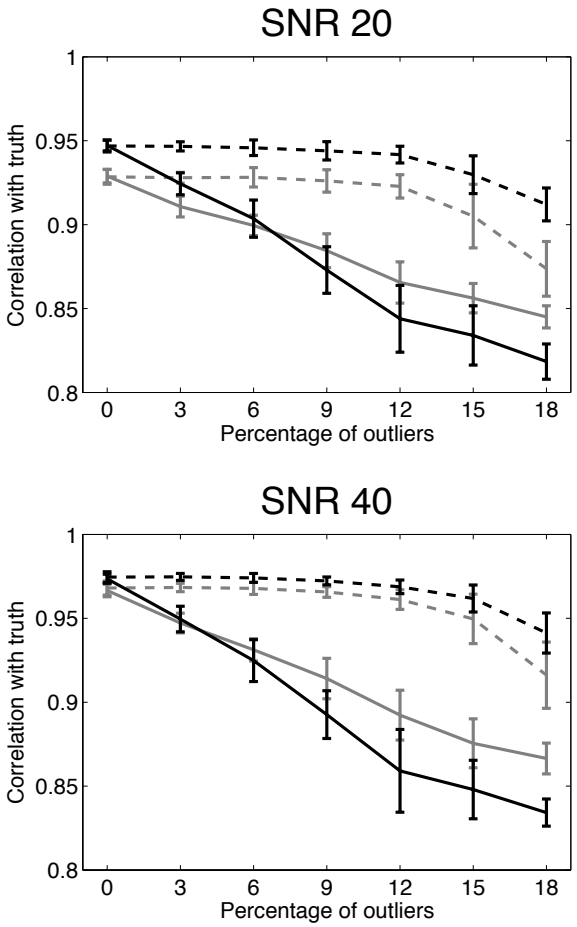

Figure 12: This figure shows the correlation between estimated and "true" FA vs frequency of outliers. The dashed lines show the situation when outlier detection and replacement was performed and the solid line when it was not. All outlier detection was performed at the 4σ level. Black lines represent tensor estimation using weighted linear fit (WLS) and grey lines represent tensor estimation with linear fit (OLS). The error bars represent one standard deviation across the ten realisations.

Subject:	WH_015	WH_017	WH_025	WH_117	WH_198	WH_212	WH_305	WH_372
Manual	28	25	94	18	27	52	83	55
OLR 3	100	150	203	120	103	140	211	156
OLR 4	66	63	166	62	60	98	112	104

Table 4: This table shows the number of outliers detected in each of the eight "bad" subjects from the Whitehall imaging data. The top row shows the number of manually detected outliers and the following two rows show the numbers detected by `eddy` with outlier thresholds of 3σ and 4σ respectively.

number of outliers for OLR 4). A detailed description of the analysis of this subject is shown in figure 13. As can be seen from the figure there is a single manually identified outlier that is not identified by OLR 4, and there are 93 outlier slices identified by both. By comparing the top panels it can be seen that the slices identified by both are those with the greatest (negative) z-score as assessed by `eddy`. This is further confirmed by the bottom panel which shows that the human observer identified all slices with z-scores greater than 20, identified most slices with a z-score greater than ten and became considerably more unreliable below ten.

An example of a slice after distortion and movement correction with and without outlier replacement is shown in figure 14. A movie of the same example can be seen in movie S5 in the Supplemental material. Examples of sagittal slices for two volumes are shown in figure 15.

6 DISCUSSION

6.1 What do we actually achieve?

The predictions that replace the outliers are linear combinations of all the non-outlier Q-space data (for that slice). Hence they contain no "new information" and the data (the outlier) is

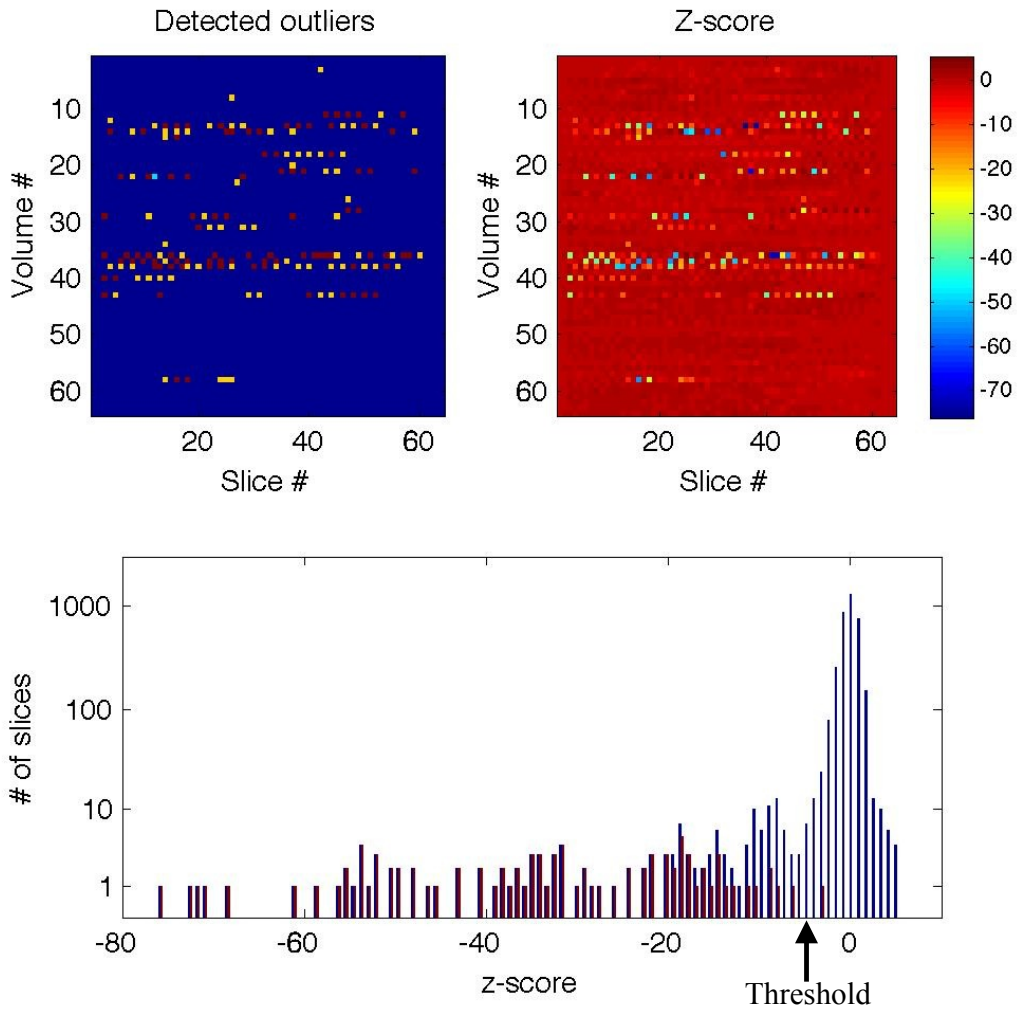

Figure 13: Detailed description of the findings in one subject. The top left panel shows the detected outliers. A dark blue color means no outlier, a light blue color means an outlier detected only manually, a yellow color an outlier detected only automatically and a red color an outlier detected by both methods. The top right panel shows the z-score for each slice as assessed by `eddy`. The bottom panel shows a histogram of the distribution of z-scores estimated by `eddy` (blue bars) and the manually detected outliers for the different z-scores (red).

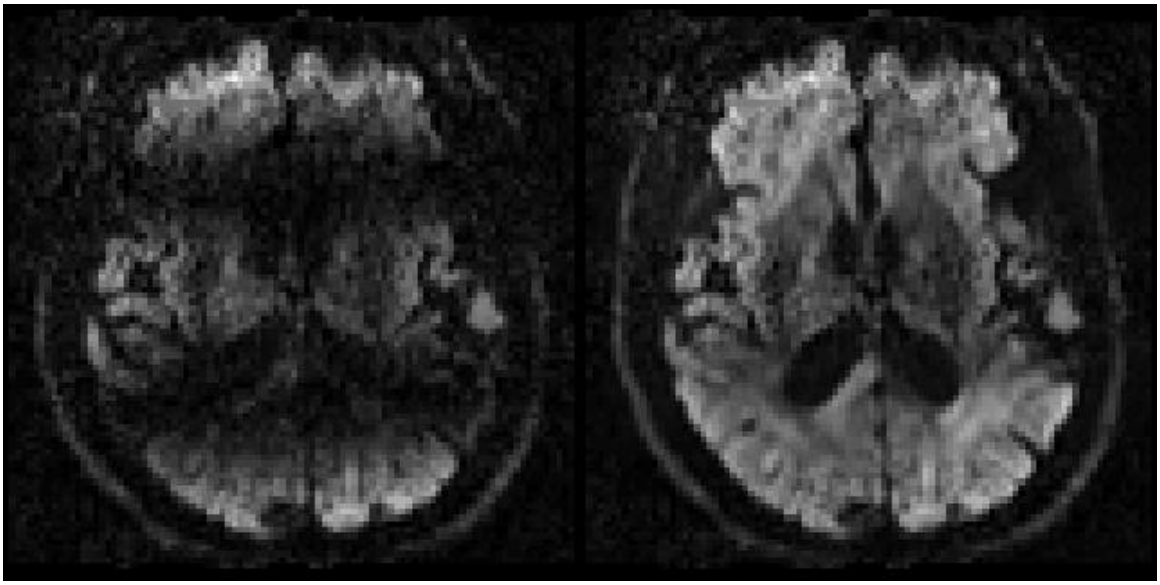

Figure 14: On the left is a slice from one dwi volume of subject WH_025, after realignment to the first volume. It shows how two outlier slices in the original volume have been rotated into wide diagonal bands in the realigned volume. On the right is the same slice where the outlier slices have been replaced in the original space prior to realignment.

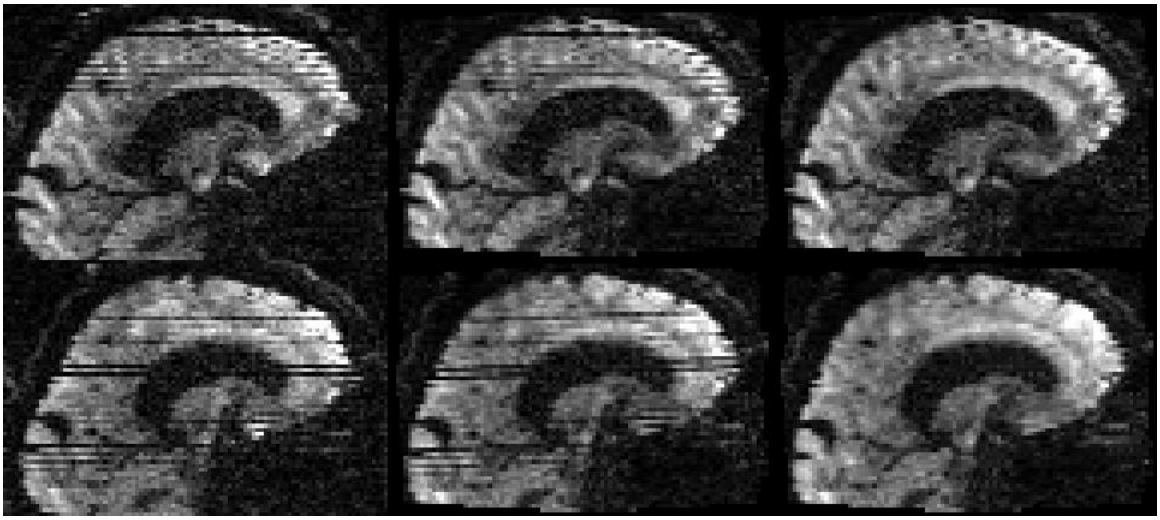

Figure 15: This figure shows a sagittal slice from two selected volumes of subject WH_025, one volume per row. The left-most panels show the situation before any correction, the middle after distortion/movement correction but *no* outlier replacement and the right-most column after correction for distortion/movement *and* outliers.

irretrievably lost. What we *can* achieve is to minimise the impact of the outlier on the subsequent diffusion analysis. The prediction that we use is formally $\hat{y}_i(\mathbf{x}_i|\mathbf{y}_{-i})$, *i.e.* our estimate of the data point y_i, associated with diffusion gradient $\mathbf{x}_i$, is *conditional* on all the other data points (denoted by $\mathbf{y}_{-i}$). In other words, the prediction is the expectation of y_i given all the other data. Due to the fact that we do not use a particular biophysical model to replace outliers, subsequent estimates of diffusion parameters from $[\mathbf{y}_{-i}\ \hat{y}_i]$ or $[\mathbf{y}_{-i}]$ will be unbiased by any inherent set of assumptions tied to a biophysical model of the diffusion process.

Given that, why not simply remove y_i instead of replacing it by $\hat{y}_i$? The reason is purely practical and related to gross subject movement. A given voxel centre in the reference space (given for example by the first b=0 volume in the diffusion dataset) is unlikely to fall on a voxel centre in the volume i that contains the outlier slice. That means that for every voxel centre we need to check where it falls in the ith volume and in particular we would need to check if it falls in the neighbourhood of the outlier slice such that we would need to use values from that slice for the interpolation. The consequence of that would be losing not just one slice, but a slab surrounding that slice. It would also preclude using more advanced interpolation models, such as spline interpolation where the entire volume affects each interpolated value (*i.e.* the interpolation "neighbourhood" is the whole volume). By instead replacing the outlier slice by "the least obtrusive" replacement we can progress with all aspects of our analysis as if there were no outlier in the first place.

It should be noted that $\hat{y}_i$ will often be associated with less uncertainty than y_i. For methods that attempt to assess the uncertainty of derived diffusion parameters (Behrens et al. (2003) and Jones (2003)) the replacement can therefore potentially lead to an underestimation of uncertainty.

It should also be noted that the accuracy of the prediction (used to replace the lost signal)

will be affected by the number of measurements in $[\mathbf{y}_{-i}]$. How critical this is will depend on the b-value. Data acquired with a higher b-value will have more high frequency information (will change more rapidly as a function of angle in Q-space) and it will therefore be more difficult to accurately predict the signal from a small number of observations. Our results demonstrate that 32 and 64 measurements are sufficient for $b = 700$ and $b = 2000$ respectively when up to 10% of the observations are missing (leaving 29 and 58 measurements in $[\mathbf{y}_{-i}]$).

6.2 Slice-wise vs voxel-wise outlier rejection

In the present paper we chose to reject and replace data on a per-slice basis. This is in contrast to voxel-wise approaches suggested by Chang et al. (2005), Chang et al. (2012), Pannek et al. (2012), Collier et al. (2015) and Tax et al. (2015). There are several reasons for our choice:

- The slice is the only real demarcation between an outlier and a non-outlier. Within a slice all the voxels have largely the same history in terms of excitation and phase-evolution and are therefore highly correlated w.r.t. k-space translation of their signal. Even in the case of pulsatile (as opposed to rigid body) movement, neighbouring voxels (within the slice) will have experienced very similar phase evolution and there is no clear border separating the outliers from the "good" voxels.

- For a given specificity, the sensitivity for detecting a single voxel as an outlier is at least an order of magnitude lower than for a slice. For a given slice the sensitivity essentially scales as $\sqrt{n_s}$, where n_s denotes the number of brain-voxels in slice s. That means that for the voxel-wise case one must either pick a lower threshold to reduce false negatives, but at the expense of inflated false positives, or keep a high threshold and have very poor sensitivity.

- With state of the art MRI scanners and multi-band sequences there is typically no paucity

of data. It is therefore more crucial to find and eliminate "bad" data in order to avoid it influencing the end result, rather than retaining a voxel that is potentially "good".

- Because we can. Our registration framework allows us to compare prediction and observation in the scan space of each volume. Previous work (Chang et al. (2005) and Chang et al. (2012), Pannek et al. (2012), Collier et al. (2015) and Tax et al. (2015)) looked for outliers after movement correction, which means that an outlier slice has potentially been distributed over several "sub-slices" (see figures 5 and 14).

We believe that the reason our method is able to combine low false positive rate and low false negative rate is precisely the use of a slice/group-wise statistic in a native (undiluted) slice-space. We have not compared our method to that suggested by Zwiers (2010), but we would expect that method to share some advantages of our method. He too assesses outliers on a slice basis (or a "patch" basis in the case of pulsatile motion) and maps the residuals back into native space. Oguz et al. (2014) goes a step further and removes the entire volume when it is found to contain outlier slices. It is not clear whether that leads to greater sensitivity as the observed slices are compared to neighbouring slices in the same volume and these are not expected to be identical in the first place.

6.3 Comparison to RESTORE

In the absence of outliers RESTORE yielded FA estimates that were further from the truth compared to simply using non-linear tensor estimation. When averaged over all brain-voxels, RESTORE on outlier free data gave the same error as no outlier detection on data with outliers (figures 7 and 8 and tables S1 and S2). These results were all obtained with the CAMINO implementation of RESTORE, but a sub-set of the data was also analysed using the TORTOISE implementation and with very similar results (presented in the Supplementary Material, section

S.2.2, figure S4).

A likely explanation for the performance on outlier free data is that any detected outlier will be a false positive, and can hence only make matters worse. As discussed before, voxelwise detection makes the trade off between false positives and negatives less advantageous (compared to slice-wise detection), meaning that there are potentially lots of false positives, which is what we observed in our simulations.

We also observed that these outliers were mainly along edges between CSF and brain, and that may explain why RESTORE fares better when calculating the correlation (with truth) only for white matter voxels.

Finally, it should be noted that when applied to single shell data RESTORE performed considerably better as can be seen in figures S2 and S3 in the supplementary material.

6.4 Impact of outliers on registration parameters

As was expected the presence of outliers had an effect on our ability to accurately estimate EC-induced distortions and subject movement. However, for the frequency of outliers used in our simulations (3%), this effect was smaller than we had expected. As can be seen in figure 6, demonstrating a typical case, the impact was limited to a small number of volumes that had a greater than average number of outliers. Even for those volumes the impact was typically limited to a few tenths of a mm increase in registration error.

Hence we conclude that the "traditional" approach of first correcting for movement and distortions followed by outlier correction is feasible, albeit less optimal than the simultaneous approach suggested in the present paper. In addition, it is not feasible to perform a *slice-wise* outlier correction following movement correction (see figures 5). In that case one would have to use a voxel-wise test, which brings pitfalls as outlined above in section 6.2.

6.5 Impact of outliers on FA and MD

As is shown in figure 11 the outliers have a significant effect on both FA and MD. For the simulations used here any given slice was affected by signal loss in 1.6 volumes on average (out of 96 dwi volumes) and this leads to a drop in correlation with the ground truth of 0.02–0.03. More outliers means a greater impact, with a drop in correlation of 0.1–0.15 when outliers constitute 18% of the slices (figure 12).

It is also shown in figures 11 and 12 that the present method can counteract the deleterious effects of outliers almost completely as long as they constitute no more than 10% of the slices.

6.5.1 Small versus large movement

It can be seen from figures 8 and 11 that the "correlation with truth" seems to be consistently a little higher for the "large movement" case compared to the "small movement" case. We believe this is related to an observation by Graham et al. (2016) that `eddy` seems to perform slightly better in terms of EC correction in the presence of some subject movement.

6.6 Relevance for group comparisons

There is great interest in and a number of methodologies for comparing FA or MD (or other diffusion derived measures) between groups (see for example Smith et al. (2006), Yushkevich et al. (2008), Maddah et al. (2011) and Raffelt et al. (2015)). Common to all of these is an implicit assumption that the quality of the input data is similar for the groups in question. That assumption is potentially violated if one of the groups happens to move more, and the result might be incidental differences such as those reported in Yendiki et al. (2014). This is not an uncommon situation for example when comparing healthy elderly subjects to those suffering from tremor or dementia (Perea et al. (2013)), healthy children to children with ADHD (Nagel

et al. (2011)) or simply different age groups (Westlye et al. (2010)). It is therefore imperative that any pre-processing is able to correct for all adverse effects of movement, and failing that at the very least detect it. We have demonstrated in previous work that `eddy` is superior to a competing method (`eddy_correct`) when it comes to estimating movement both in simulated (Graham et al. (2016)) and real data (Andersson and Sotiropoulos (2016)). In the present paper we have further demonstrated that we can almost completely correct for the signal dropout associated with subject movement. In addition `eddy` will produce a set of summary statistics reflecting total amount of movement and number of detected outliers, which could be used as regressors in the group statistical analysis.

6.7 Positive outliers

In the present paper we have defined an outlier as a slice that has, averaged across the slice, less than expected signal. This is in accordance with for example Chang et al. (2012), which revised the method in Chang et al. (2005) to only recognise signal loss. Others (Chang et al. (2005), Zwiers (2010), Pannek et al. (2012), Collier et al. (2015) and Tax et al. (2015)), in contrast, have defined outliers as residuals greater than some number of noise standard deviations regardless of direction.

Many of the causes for positive outliers, for example RF-artefacts, spiking and insufficient fat-saturation, can in principle be dealt with at the acquisition stage and be avoided altogether. In contrast, no acquisition stage remedy exists for subject movement-induced signal dropout. This is the reason we have chosen to focus only on negative outliers even if in principle our framework could be easily adapted to handle hyperintensities.

6.8 Where to next?

In the present and in an earlier paper (Andersson and Sotiropoulos (2016)) we have described methods to correct for off-resonance distortions, subject movement and movement-induced signal dropout. The problems that remain are both associated with large subject movements and are

Movement-by-susceptibility interaction: When the subject moves, especially if it involves a rotation around an axis non-parallel to the main magnetic field, the susceptibility-induced off-resonance field changes (Andersson et al. (2001)). This means that a single susceptibility-induced field is not sufficient to fully correct the distortions.

Within volume movement: The movement model use in Andersson and Sotiropoulos (2016) assumes that any volume moves "as a whole", following a rigid body model. As can be seen in figure 16 this is not necessarily true and a movement model (*i.e.* a slice-to-volume model) that takes this into account is needed (see for example Kim et al. (1999)).

The two issues are related in that they both tend to be a problem primarily in subjects who move a lot.

The first issue depends on the total amount of movement (rotation) during the scan while the second is an issue mainly for movements that are "fast" relative to the repetition time. Hence we have identified the first as the main issue in the HCP project that is characterised by young adult subjects, long scan time and short repetition time (Van Essen et al. (2013)).

In contrast the Whitehall Imaging substudy (Filippini et al. (2014)) is characterised by elderly subjects, short scan times and relatively long repetition time, and there we have identified the second issue as having greater impact. In figure 16 we demonstrate the presence of this problem in subject WH_025. As we outlined in section 3.1, the movement does not need to be very big (a rotation of 0.5 degrees was used in that example) in order to cause dropout as long as it occurs at

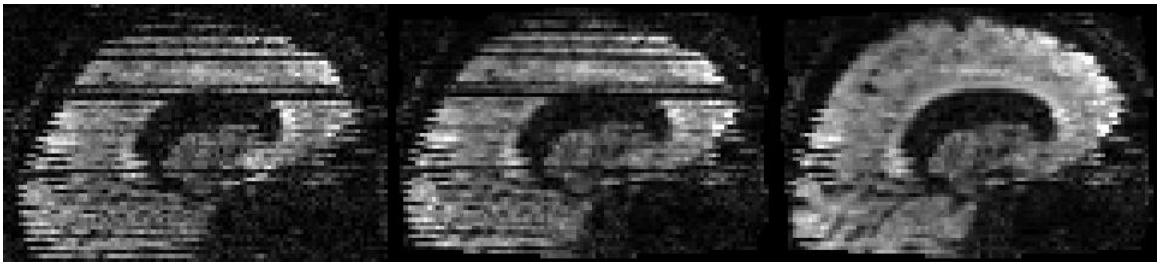

Figure 16: This figure shows a sagittal slice from one volumes of subject WH_025. The left-most panel shows the situation before any correction, the middle after distortion and movement correction but *no* outlier replacement and the right-most column after correction for distortion and movement *and* outliers. It shows a case where there is signal dropout *and* gross movement during the acquisition of the volume. Because the data was acquired with interleaved slice order this leads to the telltale "zig-zag edges" of the brain. In the right-most column it can be seen how the outlier replacement has addressed the first but not the other of these two problems.

the "wrong time" in the acquisition. But dropouts become more likely in the presence of large, sudden movements and hence they are often observed together in real data. In the present paper we have aimed to very specifically deal with the problem of outliers as caused by the mechanism described in section 3.1. Figure 16 shows that we have achieved that, but that the problem of "within volume movement" remains. We can also "correct" for the "within volume movement" problem by detecting and replacing outliers based on sum-of-squared slice differences (part of the implementation within `eddy`). However, that would imply potentially discarding valid data that should ideally be used as part of a slice-to-volume resampling strategy. Hence we have chosen not to show that and instead aim at solving the problem properly in future work.

We plan to address both "movement-by-susceptibility interaction" and "within volume movement" in future work.

6.9 Conclusion

Signal dropout is a problem for diffusion imaging, especially when acquiring data on patient populations, elderly, children *etc*. In the present paper we have shown that our strategy of joint EC, movement and outlier correction does no "harm" when applied to data without outliers and that it almost completely nullifies the deleterious effects of outliers when they are present. Our method is non-parametric and it can be applied to both single- and multi-shell dMRI data.

7 ACKNOWLEDGMENTS

We would like to thank Steve Smith and Mark Jenkinson for support and helpful discussions and Chloe Hutton for help with the English language. We would also like to thank Klaus Ebmeier for making the data from the Whitehall project available to us. Finally we gratefully acknowledge the support from the NIH Human Connectome Project (1U54MH091657-01, S.N.S and J.L.R.A.), EPSRC grant EP/L023067/1 (S.N.S.), Wellcome-Trust Strategic Award 098369/Z/12/Z (J.L.R.A.), EPSRC grant EP/L504889/1 (M.S.G.), the EPSRC Centre for Doctoral Training (EP/L016478/1, M.S.G.) and the Medical Research Council (UK) Grant MRC G1001354 (The Whitehall MRI substudy and E.Zs.).

8 REFERENCES

References

A. W. Anderson and J. C. Gore. Analysis and correction of motion artifacts in diffusion weighted imaging. *Magnetic Resonance in Medicine*, 32:379–387, 1994.

J. L. R. Andersson and S. Skare. Chapter 17: Image distortion and its correction in diffusion MRI. In D. K. Jones, editor, *Diffusion MRI: Theory, Methods, and Applications*, pages 285–302. Oxford University Press, Oxford, United Kingdom, 2011.

J. L. R. Andersson and S. N. Sotiropoulos. Non-parametric representation and prediction of single- and multi-shell diffusion-weighted MRI data using Gaussian processes. *NeuroImage*, 122:166–176, 2015. doi: 10.1016/j.neuroimage.2015.07.067.

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

J. L. R. Andersson, C. Hutton, J. Ashburner, R. Turner, and K. Friston. Modelling geometric deformations in EPI time series. *NeuroImage*, 13:903–919, 2001.

D. Atkinson, D. A. Porter, D. L. G. Hill, F. Calamante, and A. Connely. Sampling and reconstruction effects due to motion in diffusion-weighted interleaved echo planar imaging. *Magnetic Resonance in Medicine*, 44:101–109, 2000.

T. E. J. Behrens, M. W. Woolrich, M. Jenkinson, H. Johansen-Berg, R. G. Nunes, S. Clare, P. M. Matthews, J. M. Brady, and S. M. Smith. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, 50:1077–1088, 2003.

L.-C. Chang, D. K. Jones, and C. Pierpaoli. RESTORE: Robust estimation of tensors by outlier rejection. *Magnetic Resonance in Medicine*, 53:1088–1095, 2005.

L.-C. Chang, L. Walker, and C. Pierpaoli. Informed RESTORE: A method for robust estimation of diffusion tensor from low redundancy datasets in the presence of physiological noise artifacts. *Magnetic Resonance in Medicine*, 68(5):1654–1663, 2012.

Q. Collier, J. Veraart, B. Jeurissen, A. J. den Dekker, and J. Sijbers. Iterative reweighted linear least squares for accurate, fast, and robust estimation of diffusion magnetic resonance parameters. *Magnetic Resonance in Medicine*, 73(6):2174–2184, 2015.

P. A. Cook, Y. Bai, S. Nedjati-Gilani, K. K. Seunarine, M. G. Hall, G. J. Parker, and D. C. Alexander. Camino: Open-source diffusion-mri reconstruction and processing. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, volume 14, page 2759, 2006.

I. Drobnjak, D. Gavaghan, E. Süli, J. Pitt-Francis, and M. Jenkinson. Development of a functional magnetic resonance imaging simulator for modeling realistic rigid-body motion artifacts. *Magnetic Resonance in Medicine*, 56:364–380, 2006.

I. Drobnjak, G. S. Pell, and M. Jenkinson. Simulating the effects of time-varying magnetic fields with a realistic simulated scanner. *Magnetic Resonance Imaging*, 28:1014–1021, 2010.

N. Filippini, E. Zsoldos, R. Haapakoski, C. E. Sexton, A. Mahmood, C. L. Allan, A. Topiwala, V. Valkanova, E. J. Brunner, M. J. Shipley, E. Auerbach, S. Moeller, K. Uğurbil, J. Xu, E. Yacoub, J. Andersson, J. Bijsterbosch, S. Clare, L. Griffanti, A. T. Hess, M. Jenkinson, K. L. Miller, G. Salimi-Khorshidi, S. N. Sotiropoulos, N. L. Voets, S. M. Smith, J. R. Geddes, A. Singh-Manoux, C. E. Mackay, M. Kivimäki, and K. P. Ebmeier. Study protocol: The whitehall II imaging sub-study. *BMC Psychiatry*, 14:159, Jan 2014.

M. S. Graham, I. Drobnjak, and H. Zhang. Realistic simulation of artefacts in diffusion MRI for validating post-processing correction techniques. *NeuroImage*, 125:1079–1094, 2016. doi: 10.1016/j.neuroimage.2015.11.006.

R. M. Heidemann, A. Anwander, T. Feiweier, T. R. Knösche, and R. Turner. k-space and q-space: combining ultra-high spatial and angular resolution in diffusion imaging using ZOOPPA at 7 T. *NeuroImage*, 60:967–978, 2012.

H. Johansen-Berg and T. E. J. Behrens, editors. *Diffusion MRI: From Quantitative Measurement to* In-vivo *Neuroanatomy*. Academic Press, Elsevier, London, United Kingdom, 2014.

D. K. Jones. Determining and visualizing uncertainty in estimates of fiber orientation from diffusion tensor MRI. *Magnetic Resonance in Medicine*, 49(1):7–12, 2003.

D. K. Jones, editor. *Diffusion MRI: Theory, Methods, and Applications*. Oxford University Press, Oxford, United Kingdom, 2011.

D. K. Jones and P. J. Basser. "squashing peanuts and smashing pumpkins": How noise distorts diffusion-weighted MR data. *Magnetic Resonance in Medicine*, 52:979–993, 2004.

D. K. Jones, M. A. Horsfield, and A. Simmons. Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging. *Magnetic Resonance in Medicine*, 42: 515–525, 1999.

B. Kim, J. L. Boes, P. H. Bland, T. L. Chenevert, and C. R. Meyer. Motion correction in fMRI via registration of individual slices into an anatomical volume. *Magnetic Resonance in Medicine*, 41:964–972, 1999.

S. K. Krogsrud, A. M. Fjell, C. K. Tamnes, H. Grydeland, L. Mork, P. Due-Tønnessen, A. Bjørnerud, C. Sampaio-Baptista, J. Andersson, H. Johansen-Berg, and K. B. Walhovd.

Changes in white matter microstructure in the developing brain—a longitudinal diffusion tensor imaging study of children from 4 to 11 years of age. *NeuroImage*, 124:473–486, 2016.

M. Maddah, J. V. Miller, E. V. Sullivan, A. Pfefferbaum, and T. Rohlfing. Sheet-like white matter fiber tracts: representation, clustering, and quantitative analysis. *Medical Image Computing and Computer-Assisted Intervention*, 14:191–199, 2011.

J.-F. Mangin, C. Poupon, C. Clark, D. Le Bihan, and I. Bloch. Distortion correction and robust tensor estimation for MR diffusion imaging. *Medical Image Analysis*, 6(3):191–198, 2002.

S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Ugurbil. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, 63(5):1144–1153, 2010.

S. Mohammadi, C. Hutton, Z. Nagy, O. Josephs, and N. Weiskopf. Retrospective correction of physiological noise in dti using an extended tensor model and peripheral measurements. *Magnetic Resonance in Medicine*, 70:358–369, 2013.

B. J. Nagel, D. Bathula, M. Herting, C. Schmitt, C. D. Kroenke, D. Fair, and J. T. Nigg. Altered white matter microstructure in children with attention deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50:283–292, 2011.

D. G. Norris. Implications of bulk motion for diffusion-weighted imaging experiments: effects, mechanisms, and solutions. *Journal of Magnetic Resonance Imaging*, 13:486–495, 2001.

R. G. Nunes, P. Jezzard, and S. Clare. Investigations on the efficiency of cardiac-gated methods for the acquisition of diffusion-weighted images. *Journal of Magnetic Resonance*, 177:102–110, 2005.

I. Oguz, M. Farzinfar, J. Matsui, F. Budin, Z. Liu, G. Gerig, H. J. Johnson, and M. Styner. DTIPrep: quality control of diffusion-weighted images. *Frontiers in Neuroinformatics*, 8(4): 1–11, 2014.

K. Pannek, D. Raffelt, C. Bell, J. L. Mathias, and S. E. Rose. HOMOR: higher order model outlier rejection for high b-value MR diffusion data. *NeuroImage*, 63:835–842, 2012.

R. D. Perea, R. C. Rada, J. Wilson, E. D. Vidoni, J. K. Morris, K. E. Lyons, R. Pahwa, J. M. Burns, and R. A. Honea. A comparative white matter study with parkinson's disease, parkinson's disease with dementia and alzheimer's disease. *Journal of Alzheimers Disease and Parkinsonism*, 3:123–139, 2013.

C. Pierpaoli. Chapter 18: Artifacts in diffusion MRI. In D. K. Jones, editor, *Diffusion MRI: Theory, Methods, and Applications*, pages 303–318. Oxford University Press, Oxford, United Kingdom, 2011.

C. Pierpaoli, L. Walker, M. O. Irfanoglu, A. Barnett, P. Basser, L.-C. Chang, C. Koay, S. Pajevic, G. Rohde, J. Sarlls, and M. Wu. Tortoise: an integrated software package for processing of diffusion mri data. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, volume 18, page 1597, 2010.

D. A. Raffelt, R. E. Smith, G. R. Ridgway, J.-D. Tournier, D. N. Vaughan, S. Rose, R. Henderson, and A. Connelly. Connectivity-based fixel enhancement: Whole-brain statistical analysis of diffusion mri measures in the presence of crossing fibres. *NeuroImage*, 117:40–55, 2015.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, Cambridge, Massachusetts, 2006.

T. G. Reese, O. Heid, R. M. Weisskoff, and V. J. Vedeen. Reduction of eddy-current-induced

distortion in diffusion mri using a twice-refocused spin echo. *Magnetic Resonance in Medicine*, 49:177–182, 2003.

K. Setsompop, B. A. Gagoski, J. R. Polimeni, T. Witzel, V. J. Wedeen, and L. L. Wald. Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar maging with reduced g-factor penalty. *Magnetic Resonance in Medicine*, 67(5):1210–1224, 2012.

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

S. Skare and J. L. R. Andersson. On the effects of gating in diffusion imaging of the brain using single shot epi. *Magnetic Resonance Imaging*, 19:1125–1128, 2001.

S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. D. Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. D. Stefano, J. M. Brady, and P. M. Matthews. Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage*, 23(S1):208–219, 2004.

S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. Nichols, C. Mackay, K. Watkins, O. Ciccarelli, M. Cader, P. Matthews, and T. Behrens. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31:1487–1505, 2006.

S. N. Sotiropoulos, S. Jbabdi, J. Xu, J. L. Andersson, S. Moeller, E. J. Auerbach, M. F. Glasser, M. Hernandez, G. Sapiro, M. Jenkinson, D. A. Feinberg, E. Yacoub, C. Lenglet, D. C. V.

Essen, K. Ugurbil, T. E. J. Behrens, and W.-M. H. Consortium. Advances in diffusion MRI acquisition and processing in the human connectome project. *NeuroImage*, 80:125–143, 2013.

P. Storey, F. J. Frigo, H. S. R, B. J. Mock, B. D. Collick, N. Baker, J. Marmurek, and S. J. Graham. Partial k-space reconstruction in single-shot diffusion-weighted echo-planar imaging. *Magnetic Resonance in Medicine*, 57:614–619, 2007.

C. M. W. Tax, W. M. Otte, M. A. Viergever, R. M. Dijkhuizen, and A. Leemans. REKINDLE: robust extraction of kurtosis INDices with linear estimation. *Magnetic Resonance in Medicine*, 73(2):794–808, 2015.

T. P. Trouard, Y. Sabharwal, M. I. Altbach, and A. F. Gmitro. Analysis and comparison of motion-correction techniques in diffusion-weighted imaging. *Journal of Magnetic Resonance Imaging*, 6:925–935, 1996.

K. Uğurbil, J. Xu, E. J. Auerbach, S. Moeller, A. T. Vu, J. M. Duarte-Carvajalino, C. Lenglet, X. Wu, S. Schmitter, P. F. V. de Moortele, J. Strupp, G. Sapiro, F. D. Martino, D. Wang, N. Harel, M. Garwood, L. Chen, D. A. Feinberg, S. M. Smith, K. L. Miller, S. N. Sotiropoulos, S. Jbabdi, J. L. Andersson, T. E. Behrens, M. F. Glasser, D. C. V. Essen, and E. Yacoub. Pushing spatial and temporal resolution for functional and diffusion MRI in the human connectome project. *NeuroImage*, 80:80–104, 2013.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013.

A. T. Vu, E. Auerbach, C. Lenglet, S. Moeller, S. N. Sotiropoulos, S. Jbabdi, J. Andersson, E. Yacoub, and K. Ugurbil. High resolution whole brain diffusion imaging at 7 T for the Human Connectome Project. *NeuroImage*, 2015. doi: 10.1016/j.neuroimage.2015.08.004.

H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, Berlin / Heidelberg, 2003.

L. Walker, L.-C. Chang, A. Nayak, M. O. Irfanoglu, K. N. Botteron, J. McCracken, R. C. McKinstry, M. J. Rivkin, D.-J. Wang, J. Rumsey, and C. Pierpaoli. The diffusion tensor imaging (DTI) component of the NIH MRI study of normal brain development (PedsDTI). *NeuroImage*, 124B:1125–1130, 2016.

V. J. Wedeen, R. M. Weisskoff, and B. P. Poncelet. MRI signal void due to in-plane motion is all-or-none. *Magnetic Resonance in Medicine*, 32(1):116–120, 1994.

L. T. Westlye, K. B. Walhovd, A. M. Dale, A. Bjørnerud, P. Due-Tønnessen, A. Engvig, H. Grydeland, C. K. Tamnes, Y. Østby, and A. M. Fjell. Life-span changes of the human brain white matter: Diffusion tensor imaging (DTI) and volumetry. *Cerebral Cortex*, 20:2055–2068, 2010.

A. Yendiki, K. Koldewyn, S. Kakunoori, N. Kanwisher, and B. Fischl. Spurious group differences due to head motion in a diffusion MRI study. *NeuroImage*, 88:79–90, 2014.

P. A. Yushkevich, H. Zhang, T. J. Simon, and J. C. Gee. Structure-specific statistical mapping of white matter tracts. *NeuroImage*, 41:448–461, 2008.

Z. Zhou, W. Liu, J. Cui, X. Wang, D. Arias, Y. Wen, R. Bansal, X. Hao, Z. Wang, B. S. Peterson, and D. Xu. Automated artifact detection and removal for improved tensor estimation in motion-corrupted DTI data sets using the combination of local binary patterns and 2D partial least squares. *Magnetic Resonance Imaging*, 29:230–242, 2011.

M. P. Zwiers. Patching cardiac and head motion artefacts in diffusion-weighted images. *NeuroImage*, 53(2):565–575, 2010.

S Supplementary material

Listed below are text, tables and figures that were deemed superfluous to the main article.

S.1 Understanding the origin of the dropout

Below we overview the effect of bulk motion on a single-shot sequence. We assume a Stejskal-Tanner type diffusion encoding (but a similar reasoning can be made for other types of encoding, for example Reese et al. (2003)) for which the diffusion weighting is given by

$$b = \gamma^2 G^2 \delta^2 (\Delta - \delta/3) \tag{S1}$$

where γ is the gyromagnetic ratio, G and δ are the diffusion encoding gradient strength and duration (of one lobe) respectively and Δ is the time between the two gradients. Without loss of generality, let us assume that G =30 mT/m and that δ =25 ms (Δ is immaterial for this) and that the extent of the FOV is 128 voxels $\times$ 2 mm = 256 mm (which means that the centres of the end voxels are 254mm apart). These would be typical values for a "standard" diffusion weighted image acquired with a modern 3T scanner and a b-value of 1500s/mm^2. Let us also assume rectangular gradients, so that the phase accrued at location r along a gradient direction during the application of a gradient is $Gr\gamma\delta$.

Consider a column of voxels along the y-direction and imagine that the diffusion gradient is applied along the z-direction (see top row of figure S1). When the diffusion gradient is played out there will be a uniform phase accrual for all voxels along the column, *i.e.* there will be no dephasing of signal along the column. Now imagine that in the period between the two lobes of the diffusion gradient the subject moves such that there is a rotation of 0.5 degrees around the x-axis (*i.e.* an axis orthogonal to both the column and the diffusion gradient). This means that when the second lobe of the diffusion gradient is played out the centres of the ultimate voxels

along that column will be a distance 254 mm $\times \sin .5^\circ \approx 2.2$ mm apart and that the voxels along that column will accrue a relative phase that is proportional to their location. The phase difference between the ultimate voxels is given by $Gr\gamma\delta$, *i.e.* 0.03 T/m $\times$ 0.254 m $\times \sin .5^\circ \times$ $267.5{\cdot}10^6$ rad s^{-1} T^{-1} $\times$ 0.025 s $\approx$ 445 radians. A linear phase of 2π in the image domain corresponds to a one-pixel translation in k-space, which means that the 445 radians correspond to a 71 pixel shift of the signal along the "y-direction" in k-space. Those 71 voxels will be enough to move the centre, and most of the power, of the signal outside the sampled window, leading to an almost total loss of signal.

Next we want to demonstrate how the exact same movement can cause a very different (negligible) dropout for a different direction diffusion gradient. Everything is the same as above, *except* that the diffusion gradient is applied along the y-direction (see bottom row of figure S1). For this case the voxels along the column will accrue a relative phase that is proportional to their location along the column such that the phase difference between the ultimate voxels is given by 0.03 T/m $\times$ 0.254 m $\times$ $267.5{\cdot}10^6$ rad s^{-1} T^{-1} $\times$ 0.025 s $\approx$ 50959 radians ($Gr\gamma\delta$). As before, in the period between the two lobes of the diffusion gradient the subject moves resulting in a rotation of 0.5 degrees around an axis orthogonal to the diffusion gradient. That means that the distance (projected onto the diffusion gradient) between the points that previously corresponded to the centres of the ultimate voxels is now "only" 254 mm $\times \cos 0.5^\circ \approx 253.99$ mm. Hence, between the ultimate voxels, the second lobe will rewind $(253.99/254) \times 5 \cdot 10^4 \approx 50957$ radians, leaving 2 radians linear phase along the direction of the diffusion gradient. That means that in this case the same movement results in a 1/3 pixel translation in k-space, which would cause a trivial or no loss of signal.

A subject movement in the form of a translation in the period between the two lobes would cause a constant (over space) phase which would not cause any signal loss.

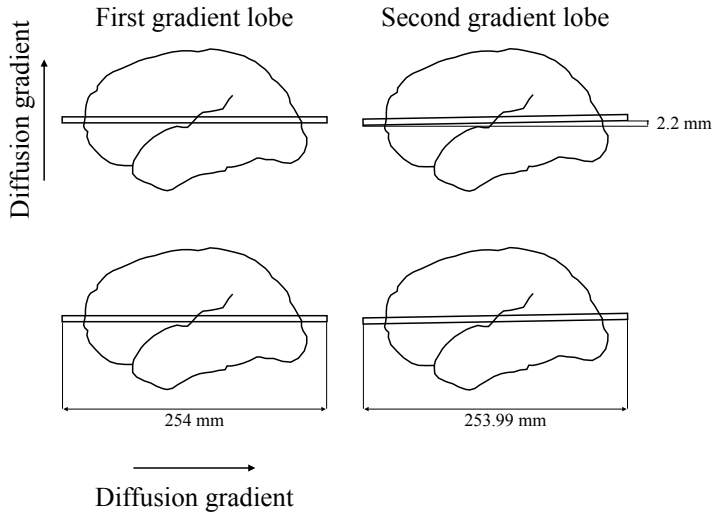

Figure S1: The left column shows the situation during the first diffusion gradient lobe and the right during the second gradient lobe (after the re-focusing pulse in an ST scheme). The top row shows the case where the diffusion gradient is perpendicular to the excited plane and the bottom row when it is parallel to the plane. In both cases there is a 0.5° rotation around the subjects left-right axis (the axis perpendicular to the sagittal view in the figure). The imaging parameters are assumed to be those specified in the main text. In the top row the first diffusion gradient causes no phase accrual along the subjects posterior-anterior axis. However, during the second gradient the posterior part experiences a stronger field than the anterior part and a linear phase of ≈ 445 radians is introduced between the end voxels. In the bottom row the first gradient results in a linear phase of ≈ 50959 radians as the posterior part experiences a stronger field. During the second gradient the distance between the extreme voxels is slightly smaller (≈ 0.01mm) so the re-phasing is not complete and ≈ 2 radians are left. The top case leads to a translation of most of the signal outside the k-space window, and hence severe dropout, whereas the bottom case leads to a trivial translation of $\approx 1/3$ of a "k-space pixel".

SNR:	20		
Data:	Without Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9638 ± .0016	.9519 ± .0020	.9640 ± .0012
White matter	.9442 ± .0038	.9432 ± .0039	.9444 ± .0030
Data:	With Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9564 ± .0030	.9458 ± .0029	.9636 ± .0027
White matter	.9345 ± .0052	.9369 ± .0050	.9438 ± .0059
SNR:	40		
Data:	Without Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9826 ± .0012	.9747 ± .0012	.9831 ± .0016
White matter	.9702 ± .0030	.9692 ± .0030	.9715 ± .0041
Data:	With Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9754 ± .0028	.9686 ± .0025	.9822 ± .0013
White matter	.9615 ± .0050	.9657 ± .0044	.9700 ± .0031

Table S1: Correlation between "ground truth" and estimated FA in the absence and presence of outliers in the data. These results pertain to the simulations with movement levels commensurate with a "good" subject.

S.2 Comparison to RESTORE

Tables S1 and S2 list the results that are the basis for figure 8 in the main text.

S.2.1 Separate analysis of low and high b-value shells

The results presented in the tables above and figures 7 and 8 in the main text are based on both shells ($b = 700$ and $b = 2000$) of the simulated data. We also performed an analysis where the data was divided up into two sets consisting of $4b = 0 + 32b = 700$ volumes and $8b = 0 + 64b = 2000$ volumes respectively. Results from this analysis are presented in figures S2 and S3.

A couple of findings are worth pointing out in figures S2 and S3.

- RESTORE performs better on single shell than on multi shell data (compare figures S2

SNR:	20		
Data:	Without Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9691 ± .0007	.9598 ± .0011	.9691 ± .0007
White matter	.9559 ± .0016	.9545 ± .0017	.9559 ± .0018
Data:	With Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9604 ± .0025	.9512 ± .0022	.9681 ± .0013
White matter	.9451 ± .0034	.9475 ± .0029	.9536 ± .0031
SNR:	40		
Data:	Without Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9860 ± .0011	.9804 ± .0012	.9859 ± .0010
White matter	.9786 ± .0024	.9774 ± .0024	.9781 ± .0021
Data:	With Outliers		
Analysis:	eddy	eddy+RESTORE	eddy with OLR
Whole brain	.9775 ± .0025	.9719 ± .0018	.9853 ± .0008
White matter	.9684 ± .0033	.9728 ± .0025	.9773 ± .0020

Table S2: Correlation between "ground truth" and estimated FA in the absence and presence of outliers in the data. These results pertain to the simulations with movement levels commensurate with a "bad" subject.

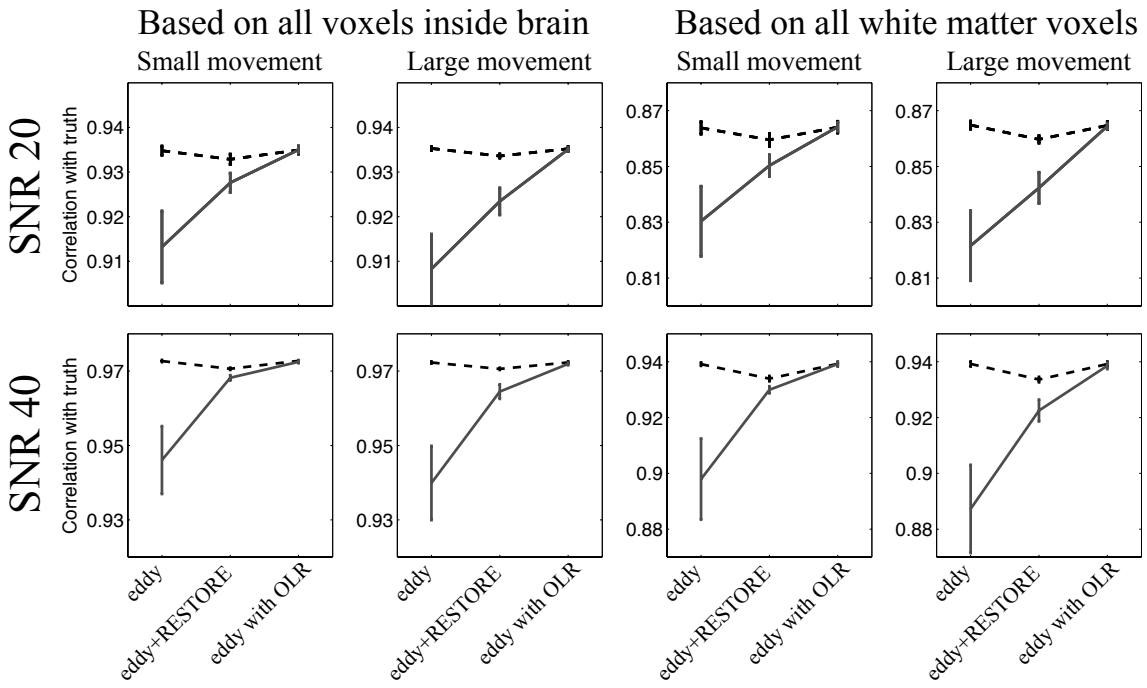

Figure S2: This figure, similarly to figure 8 in the main text, shows the voxelwise correlation between estimated FA and "ground truth" for the simulated data. The difference compared to figure 8 is that here only the data from the low-b shell ($b = 700$) was used. Each point represents the correlation across all brain-voxels (left two columns) or across all white matter voxels (right two columns) averaged over ten realisations. The error bars represent $\pm$ one standard deviation across those ten realisations. The dashed black lines represent the cases where there were no outliers in the data, and the solid grey lines where there were outliers.

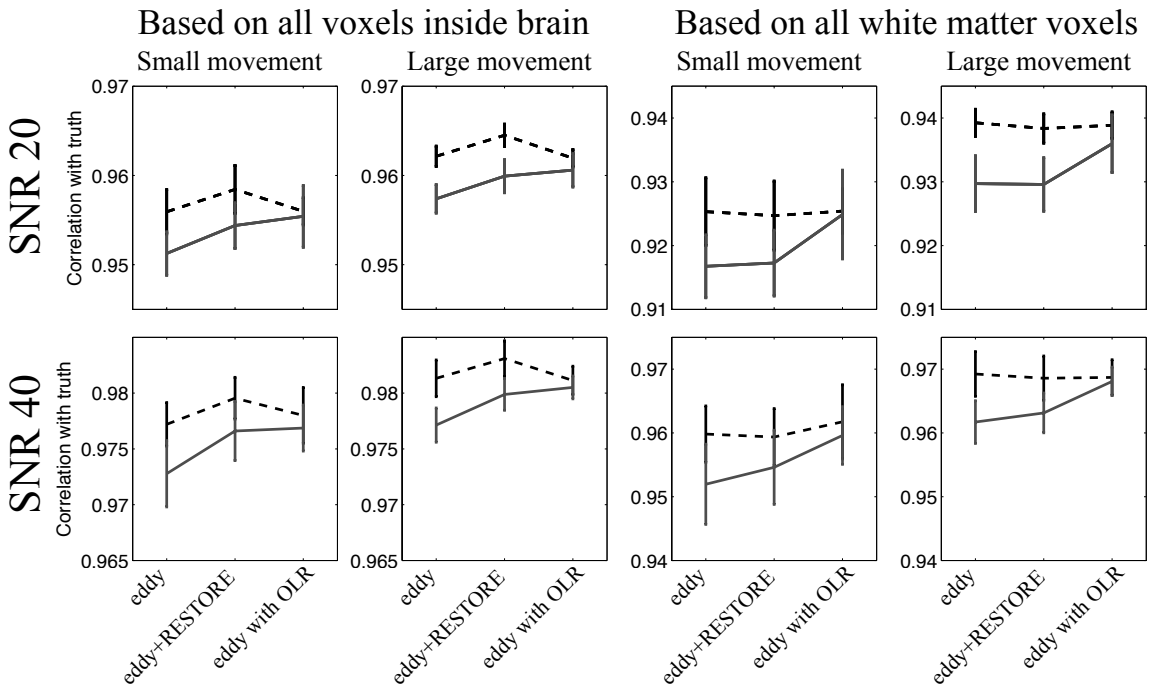

Figure S3: This figure, similarly to figure 8 in the main text, shows the voxelwise correlation between estimated FA and "ground truth" for the simulated data. The difference compared to figure 8 is that here only the data from the high-b shell ($b = 2000$) was used. Each point represents the correlation across all brain-voxels (left two columns) or across all white matter voxels (right two columns) averaged over ten realisations. The error bars represent $\pm$ one standard deviation across those ten realisations. The dashed black lines represent the cases where there were no outliers in the data, and the solid grey lines where there were outliers.

and S3 to figure 8 in the main text.

- When assessed across all voxels the deleterious effects of outliers is greater for low than for high b-value data (compare figures S2 and S3). In retrospect this is not really surprising since for a given diffusion direction the majority of voxels have little signal left at a b-value of 2000. However, note that it is likely that for the subset of voxels with largely preserved signal for a given diffusion gradient the outliers will still have a significant impact.

- In the presence of outliers `eddy` with OLR still performs better than `eddy`+RESTORE for all cases.

- When assessed for all brain-voxels it seems that RESTORE improves the results for the high b-values data *even* when there are no outliers in the data (left two columns of figure S3). The outliers were mainly detected along CSF-tissue boundaries. Hence a tentative explanation for this is that RESTORE excluded data points affected by small registration errors that has the greatest effect where there are strong gradients in the images.

S.2.2 Comparison to TORTOISE implementation of RESTORE and iRESTORE

All the comparisons above were performed using the CAMINO implementation of RESTORE. In order to ensure that the results were not specific to one particular implementation we also performed a comparison to the RESTORE and iRESTORE (Chang et al. (2012)) as implemented in TORTOISE (Pierpaoli et al. (2010)). This comparison was performed on a sub-set of the simulations and the results are presented in figure S4.

Comparing this figure to figure 8 it can be seen that

- The behaviour of TORTOISE RESTORE is very similar to that of CAMINO RESTORE.

- Compared to RESTORE, iRESTORE causes less harm to data with no outliers. This is

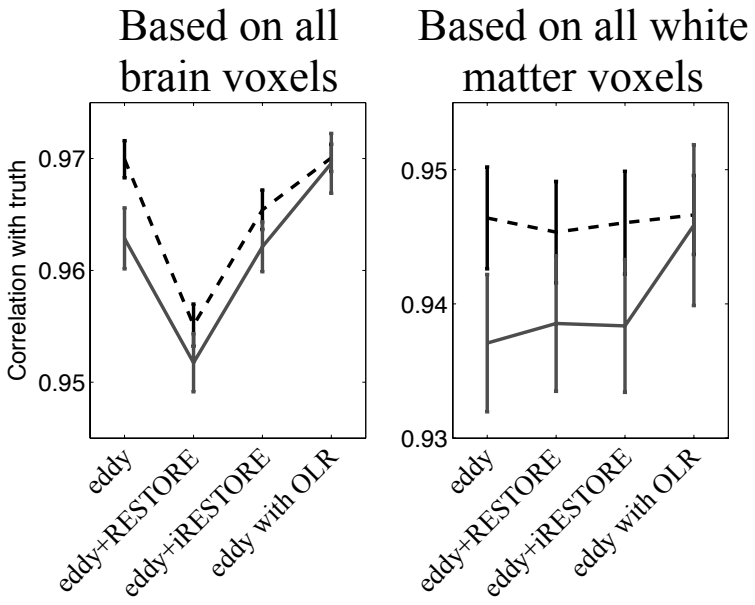

Figure S4: This figue was based on the "small movement" simulation with an SNR of 20. Hence it corresponds to panels 1 and 3 in the top row of figure 8 in the main part of the paper. The dashed black lines corresponds to simulations without any outliers and the grey solid lines with outliers. The vertical bars represent ± one standard deviation across the ten realisations of the simulations.

to be expected as one would expect iRESTORE to reduce the number of false positives by 50%.

- iRESTORE performs approximately as RESTORE with respect to improving matters in the presence of outliers.

S.3 The effect on derived diffusion parameters

Detailed results for the effect of outliers on derived diffusion parameters (FA and MD) are shown in tables S3, S4, S5 and S6. A subset of these results form the basis for figure 11 in the main text.

S.4 Movies of subject WH_025

As described in the main text, the manual assessment found 94 outliers in the "worst" subject, while the automatic detection at a threshold of 4σ found 166. The movie S5 extends figure 14 in the main text to four adjacent slices and also to flip between the corrected and uncorrected images

The movie S6 extends figure 15 in the main text to show all diffusion weighted volumes for subject WH_025.

SNR:	20				
Data:	Without Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9277 ± .0050	.9276 ± .0055	.9285 ± .0044	.9273 ± .0059	.9289 ± .0040
Weighted (WLS)	.9459 ± .0042	.9462 ± .0043	.9469 ± .0036	.9457 ± .0047	.9472 ± .0032
Non-linear (NLS)	.9436 ± .0043	.9438 ± .0045	.9446 ± .0037	.9434 ± .0049	.9448 ± .0032
Data:	With Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9272 ± .0057	.9283 ± .0055	.9268 ± .0059	.9274 ± .0062	.9122 ± .0052
Weighted (WLS)	.9451 ± .0051	.9462 ± .0048	.9448 ± .0057	.9455 ± .0058	.9254 ± .0057
Non-linear (NLS)	.9430 ± .0054	.9441 ± .0050	.9427 ± .0057	.9434 ± .0058	.9361 ± .0041
SNR:	40				
Data:	Without Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9685 ± .0043	.9676 ± .0044	.9680 ± .0042	.9672 ± .0037	.9666 ± .0038
Weighted (WLS)	.9749 ± .0032	.9741 ± .0033	.9745 ± .0032	.9739 ± .0028	.9736 ± .0027
Non-linear (NLS)	.9721 ± .0035	.9714 ± .0036	.9717 ± .0035	.9711 ± .0030	.9707 ± .0030
Data:	With Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9655 ± .0038	.9665 ± .0042	.9652 ± .0045	.9644 ± .0050	.9467 ± .0075
Weighted (WLS)	.9727 ± .0029	.9734 ± .0032	.9724 ± .0035	.9717 ± .0039	.9492 ± .0080
Non-linear (NLS)	.9699 ± .0031	.9707 ± .0034	.9696 ± .0037	.9688 ± .0041	.9611 ± .0041

Table S3: Correlation between "ground truth" and estimated FA in the absence and presence of outliers in the data. For all cases `eddy` was run either with outlier detection and replacement at a specified level (outliers defined as 3, 3.5, 4 or 4.5 σ from mean) or without outlier detection (∞ σ). These results pertain to the simulations with movement levels commensurate with a "good" subject.

SNR:	20				
Data:	Without Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9410 ± .0026	.9406 ± .0018	.9415 ± .0023	.9412 ± .0022	.9414 ± .0020
Weighted (WLS)	.9577 ± .0021	.9574 ± .0014	.9581 ± .0017	.9578 ± .0017	.9579 ± .0016
Non-linear (NLS)	.9560 ± .0020	.9558 ± .0015	.9564 ± .0016	.9562 ± .0015	.9562 ± .0015
Data:	With Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9400 ± .0035	.9394 ± .0023	.9380 ± .0038	.9389 ± .0038	.9248 ± .0052
Weighted (WLS)	.9566 ± .0025	.9564 ± .0015	.9551 ± .0028	.9558 ± .0027	.9344 ± .0062
Non-linear (NLS)	.9546 ± .0025	.9543 ± .0013	.9530 ± .0028	.9537 ± .0027	.9453 ± .0037
SNR:	40				
Data:	Without Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9741 ± .0034	.9743 ± .0039	.9746 ± .0033	.9748 ± .0040	.9749 ± .0039
Weighted (WLS)	.9803 ± .0027	.9806 ± .0029	.9808 ± .0024	.9809 ± .0031	.9810 ± .0031
Non-linear (NLS)	.9778 ± .0027	.9780 ± .0028	.9782 ± .0024	.9783 ± .0031	.9784 ± .0031
Data:	With Outliers				
No. of σ:	3	3.5	4	4.5	∞
Linear (OLS)	.9737 ± .0037	.9740 ± .0032	.9737 ± .0024	.9739 ± .0025	.9568 ± .0055
Weighted (WLS)	.9797 ± .0030	.9801 ± .0027	.9797 ± .0019	.9799 ± .0020	.9562 ± .0070
Non-linear (NLS)	.9772 ± .0030	.9776 ± .0028	.9772 ± .0020	.9774 ± .0021	.9678 ± .0040

Table S4: Correlation between "ground truth" and estimated FA in the absence and presence of outliers in the data. For all cases `eddy` was run either with outlier detection and replacement at a specified level (outliers defined as 3, 3.5, 4 or 4.5 σ from mean) or without outlier detection (∞ σ). These results pertain to the simulations with movement levels commensurate with a "bad" subject.

SNR:			20		
Data:			Without Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.8660 ± .0034	.8656 ± .0035	.8664 ± .0024	.8653 ± .0039	.8667 ± .0022
Weighted	.8654 ± .0037	.8666 ± .0025	.8670 ± .0018	.8655 ± .0035	.8665 ± .0029
Non-linear	.8652 ± .0022	.8655 ± .0019	.8659 ± .0016	.8651 ± .0022	.8660 ± .0014
Data:			With Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.8644 ± .0032	.8648 ± .0026	.8642 ± .0035	.8648 ± .0041	.8430 ± .0048
Weighted	.8637 ± .0049	.8647 ± .0038	.8646 ± .0038	.8646 ± .0056	.8557 ± .0039
Non-linear	.8644 ± .0028	.8649 ± .0021	.8642 ± .0029	.8647 ± .0030	.8628 ± .0022
SNR:			40		
Data:			Without Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.9128 ± .0022	.9121 ± .0027	.9121 ± .0024	.9117 ± .0023	.9115 ± .0025
Weighted	.8838 ± .0069	.8824 ± .0068	.8825 ± .0068	.8810 ± .0102	.8812 ± .0101
Non-linear	.8852 ± .0015	.8847 ± .0012	.8848 ± .0016	.8845 ± .0016	.8844 ± .0015
Data:			With Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.9102 ± .0023	.9107 ± .0028	.9101 ± .0019	.9096 ± .0028	.8819 ± .0098
Weighted	.8842 ± .0038	.8848 ± .0037	.8840 ± .0029	.8835 ± .0036	.8735 ± .0040
Non-linear	.8838 ± .0017	.8842 ± .0017	.8836 ± .0016	.8834 ± .0019	.8812 ± .0015

Table S5: Correlation between "ground truth" and estimated MD in the absence and presence of outliers in the data. For all cases `eddy` was run either with outlier detection and replacement at a specified level (outliers defined as 3, 3.5, 4 or 4.5 σ from mean) or without outlier detection (∞ σ). These results pertain to the simulations with movement levels commensurate with a "good" subject.

SNR:			20		
Data:			Without Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.8678 ± .0036	.8677 ± .0026	.8684 ± .0028	.8683 ± .0027	.8685 ± .0023
Weighted	.8652 ± .0021	.8651 ± .0014	.8655 ± .0018	.8651 ± .0019	.8653 ± .0019
Non-linear	.8642 ± .0010	.8642 ± .0007	.8645 ± .0008	.8644 ± .0008	.8644 ± .0008
Data:			With Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.8656 ± .0028	.8650 ± .0020	.8642 ± .0033	.8648 ± .0034	.8450 ± .0047
Weighted	.8629 ± .0027	.8628 ± .0019	.8619 ± .0026	.8618 ± .0036	.8542 ± .0035
Non-linear	.8633 ± .0013	.8632 ± .0009	.8627 ± .0014	.8629 ± .0014	.8606 ± .0017
SNR:			40		
Data:			Without Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.9090 ± .0031	.9087 ± .0036	.9088 ± .0035	.9093 ± .0029	.9093 ± .0029
Weighted	.8842 ± .0027	.8822 ± .0075	.8823 ± .0075	.8815 ± .0102	.8817 ± .0103
Non-linear	.8821 ± .0013	.8822 ± .0015	.8822 ± .0013	.8824 ± .0015	.8823 ± .0014
Data:			With Outliers		
No. of σ:	3	3.5	4	4.5	∞
Linear	.9079 ± .0028	.9079 ± .0027	.9079 ± .0021	.9080 ± .0020	.8834 ± .0074
Weighted	.8830 ± .0038	.8839 ± .0020	.8836 ± .0012	.8830 ± .0030	.8738 ± .0041
Non-linear	.8813 ± .0015	.8816 ± .0016	.8812 ± .0010	.8814 ± .0009	.8784 ± .0019

Table S6: Correlation between "ground truth" and estimated MD in the absence and presence of outliers in the data. For all cases `eddy` was run either with outlier detection and replacement at a specified level (outliers defined as 3, 3.5, 4 or 4.5 σ from mean) or without outlier detection (∞ σ). These results pertain to the simulations with movement levels commensurate with a "bad" subject.

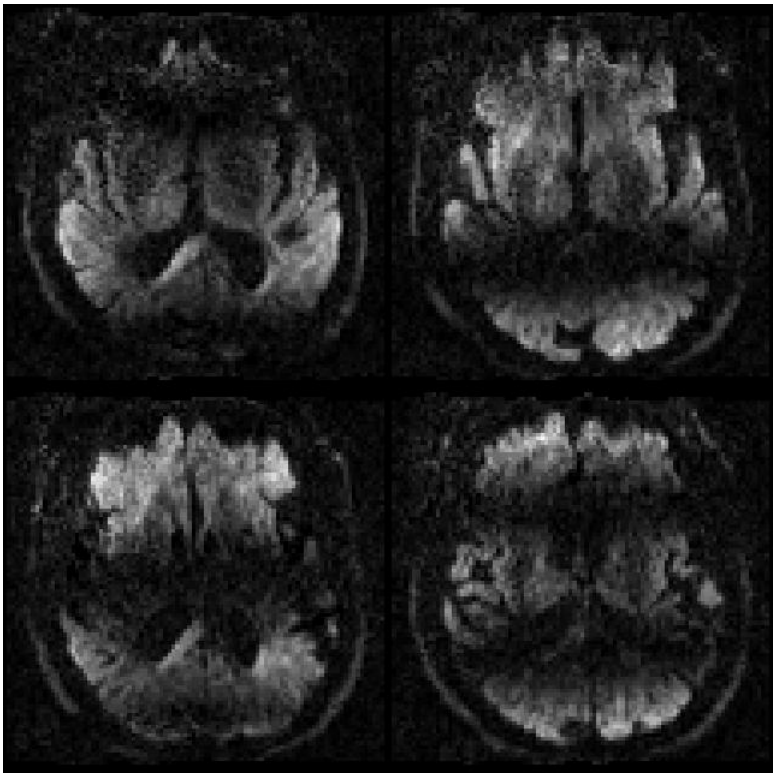

Figure S5: This is a still from the movie Movie_S1.gif. The movie extends figure 14 in the main text. It shows four contiguous transversal slices in one dwi volume from subject WH_025. The volume has been rotated into the space of the first volume of the dataset, resulting in outlier slices being transformed into wide diagonal bands in the realigned volume. The view alternates between the case with no outlier replacement and the case where outlier slices were replaced in the original volume prior to realignment.

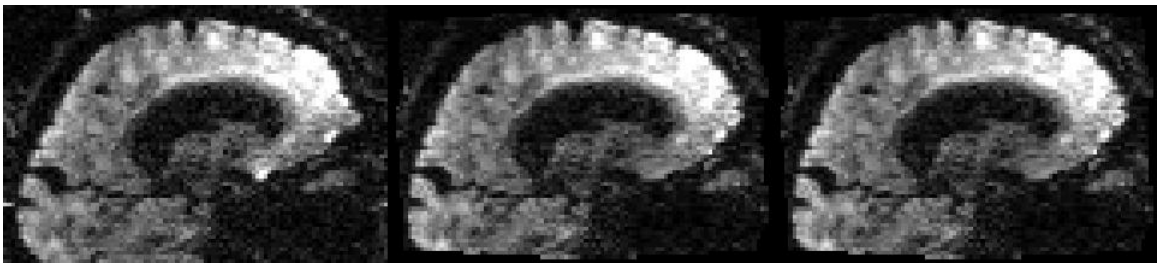

Figure S6: This is a still from the movie Movie_S2.gif. The movie extends figure 14 in the main text. It steps through all diffusion weighted volumes for subject WH_025, displaying the same sagittal slice for all. The left-most panel shows the slice prior to any correction, the middle panel shows the slice after correction for distortions and movement and the right-most panel after correction for distortions, movement and outliers.