Objective QC for diffusion MRI data: artefact detection using normative modelling

Ramona Cirstian[1,2], Natalie J. Forde[1,2], Jesper L.R. Andersson[3], Stamatios N. Sotiropoulos[3,4,5], Christian F. Beckmann[1,2,3], Andre F. Marquand[1,2,6]

1. Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
2. Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, the Netherlands
3. Wellcome Centre for Integrative Neuroimaging - Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), University of Oxford, UK
4. Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham, UK
5. National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre, Queens Medical Centre, Nottingham, UK
6. Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK

Key words: diffusion MRI, normative modelling, artefacts, quality control, Eddy QC

Contact: Ramona Cirstian (ramona.cirstian@donders.ru.nl)

Abstract

Diffusion MRI is a neuroimaging modality used to evaluate brain structure at a microscopic level and can be exploited to map white matter fiber bundles in the brain. One common issue is the presence of artefacts such as acquisition artefacts (e.g. ringing), physiological artefacts (cardiac pulsation, breathing, motion), distortions (EPI, eddy currents) or post-processing (e.g. blurring, registration errors). These may lead to problems in downstream processing and can bias subsequent analyses. In this work we use normative modeling to create a semi-automated pipeline for detecting diffusion imaging artefacts and errors by modeling 24 image derived phenotypes using the UK Biobank dataset. We used several parameter types with different modelling complexities (DTI model: FA and MD; NODDI model: ICVF and ISOVF). Likewise, we selected several tracts of various sizes and modelling difficulties (corpus callosum, corticospinal tract, uncinate fasciculus and fornix). Our method is compared to two traditional quality control modalities: a visual quality control protocol performed on 500 subjects and quantitative quality control using other automated pipelines. The normative modelling framework proves to be comprehensive and efficient in detecting diffusion imaging artefacts arising from various sources as well as outliers resulting from inaccurate post-processing (i.e., erroneous registrations). This is an important contribution by virtue of this methods' ability to identify the two problem sources (i) image artefacts and (ii) post-processing errors, which subsequently allows for a better understanding of our data and informs on inclusion/exclusion criteria of participants.

Introduction

Diffusion MRI (dMRI) is a neuroimaging modality frequently used to study the configuration of white matter in the brain. Diffusion refers to the molecular mobility of water molecules in biological tissue which can be measured in terms of its anisotropy levels (i.e. how asymmetric the diffusion pattern is within tissue). Due to the organization of white matter in bundles of myelinated axonal fibers, anisotropy can be exploited by dMRI to map the microscopic details of tissue architecture [1]. Although dMRI is a great tool for investigating in vivo structural organisation in the human brain, it is not without its difficulties. The resolution of a dMRI scan is typically much lower than a regular T1-weighted anatomical scan as well as being more predisposed to the appearance of artefacts. One of the main reasons artefacts arise is due to the increased sensitivity to off-resonance fields of the echo-planar imaging technique used for data acquisition. Likewise, the dynamic nature of collecting multiple volumes during a dMRI scan raises susceptibility to subject movement. [2]

Several biophysical models can be applied to dMRI data to estimate tissue microstructure, here we will briefly mention two of the most used models. The diffusion tensor imaging (DTI) model allows a full characterization of molecular mobility variation in space according to direction. Two of the most widely used DTI parameters are fractional anisotropy (FA) and mean diffusivity (MD).[3] Neurite orientation dispersion and density imaging (NODDI) is another popular dMRI model. NODDI produces neurite density and orientation dispersion estimates which yield to more specific markers of brain tissue microstructure than standard indices than DTI. [4]

Artefacts in dMRI scans can have the following causes and effects. Susceptibility distortions are caused by the off-resonance field created by the human head during acquisition. The water, bone and sinuses in the head all have different magnetic susceptibilities which interfere with the gradient field along the direction of the phase encoding. The artefact appears as a compression or stretching of the image, particularly noticeable in the frontal lobe. Eddy current distortions are caused by currents generated in the conducting surfaces of the scanner. The rapid gradient switching during the diffusion encoding induces eddy currents which in turn generate small magnetic fields. The interference on the magnetic gradient by these currents results in the translation of the signal in the phase-encoding direction. Artefacts created by eddy currents are often characterized by the sheared appearance of the brain in the resulting image. [2] Artefacts generated by subject movement during acquisition are especially prevalent in dMRI data. Movement can create distortions in different ways: by creating the appearance of blur, ghosting or a striped pattern (signal dropout). Furthermore, the head of the subject may be placed out of the field of view (FOV) in extreme cases from the beginning of the acquisition. In such cases, a part of the brain is missing from the scan.

The presence of artefacts in the dMRI scans may lead to problems in downstream processing and can bias subsequent analyses. For this reason, it is preferable that the artefacts are removed from the dMRI dataset before analyses. A common method for detection and removal of artefacts is performed manually by an expert in diffusion imaging. [5] Generally, a manual inspection protocol involves the visual assessment of each DWI volume pertaining to every subject. This is followed by the assignment of labels towards each artefact-affected volume (consisting of the artefact type and artefact severity score) which play a role in the decision for volume removal or subject exclusion. The labeling process requires a great deal of attention and expertise, and it is ultimately subjective in nature due to the high variation in artefact appearance. For this reason, manual inspection of large datasets is not feasible and increases the risk of either discarding valuable data or including artefactual scans into a study.

Many automated quality control (QC) tools are part of a postprocessing pipeline which are often used for detecting and correcting artefacts in larger datasets. Most of the times these pipelines base their algorithm on the exclusion of artefactual data prior further processing. The removal of data can be made at different levels of the image processing. For example, RESTORE [6] is a method for improving

tensor estimation on a voxel-by-voxel basis in the presence of artifactual data points. The algorithm detects artefactual voxels by computing an initial tensor model using non-linear least squares and evaluating the residuals. PATCH [7] is a tool for the detection and correction of motion artefacts at the slice or patch level. It uses regional and more global (slice-wise) information to detect artefacts which improves the algorithms robustness and sensitivity. At volume level, DTIprep [8] can detect and correct artefacts based on evaluating and comparing the entropy values of artefact-free scans with suspicious scans (lower values of entropy). A common issue with this type of approach is the removal of too much data, which in turn may lead to a poorly estimated model or simply data loss.

Alternatively, there are pipelines which are based on estimating a desired outcome (e.g., contrast level of the slice). For example, FSL EDDY [9,10] is a tool which retrospectively estimates various artefact types (eddy currents, susceptibility, movement, and slice dropout) by registering the individual volumes to a model free prediction of what each volume should look like. Within the same framework, EDDY QC [11] is a tool for generating qualitative single subject and group reports, summarizing a number of objective QC parameters which are acquired based on the FSL EDDY output.

Recently, deep learning-based methods for quality control of neuroimaging data have gained popularity. In the field of diffusion imaging, a few notable works emerge. The QC Automator [12] is a method based on convolutional neural networks for automated classification of dMRI volumes and uses transfer learning for the classification of axial and sagittal slices. Furthermore, 3D-QCNet [13] is a more recent algorithm which improves upon the QC Automator principle by creating a deep learning pipeline that can detect dMRI artefacts three-dimensionally without requiring manually labeled data.

As we have seen, there are extensive preprocessing pipelines designed to minimize the effect of artefacts and correct the images after the acquisition, however, some artefacts remain. After the preprocessing stage, the data is further processed by applying different models, registration and other processing steps where subsequent issues can also arise. Due to the many types of artefacts which can be affecting diffusion data, the detection of errors within the images is either time consuming, in the case of manually labeled scans, or prone to inaccuracies in the case of automated pipelines (e.g. classifiers). Many of these artefacts may be quite rare, which makes it difficult to obtain a sufficiently large, labelled dataset for training automated QC methods. Another issue with the existing pipelines arises from the many correction steps that are applied during the processing. While some artefacts may be 'corrected', some are incompletely corrected or missed. Furthermore, approaches based on examining only the data may miss artefacts that occur during downstream processing, for example, normalisation errors, which can be more serious in diffusion data than in other modalities due to the precise space localized of the brain structures.

In this paper, we propose a method for evaluating and understanding the quality of dMRI data at the subject level using normative modeling. Normative modelling [14] is an innovative method used to model biological and behavioral variation across a study population and can be used to make statistical inferences at an individual level. This is achieved by mapping a response variable (e.g., neuroimaging derived phenotypes) to a covariate (e.g., age) in a similar way growth charts are used in pediatric medicine to map the height or weight of children to their age. We aim to detect subjects with either poor quality data and/or with processing problems as extreme outliers from a normative model that captures population variation within each image derived phenotype (IDP). Crucially: (i) this does not require us to label artefacts in advance, nor (ii) specify what the consequences of different types of artefacts are on the derived phenotype, and (iii) allows immediate identification of potentially problematic scans from large datasets without labour-intensive manual screening. For this purpose, we use the UK Biobank dataset [15,16] which is one of the largest biomedical databases and research resources (currently) containing over 60 000 subjects with available diffusion data as well as hundreds of diffusion image derived phenotypes (IDPs).

Methods

Dataset

We used a subset of the UK Biobank dataset containing 23158 subjects with available dMRI data, from the 2020 data release. For a brief description of the acquisition and pre-processing steps, data were acquired using a Siemens Skyra 3T and has a duration of 7 minutes. It contains 5 x b0 scans, 50 x b1000 and 50 x b2000 with gradient timings δ = 21.4 ms, Δ = 45.5 ms. The resolution is 2x2x2 mm and the field of view is a 104x104x72 matrix. The data were corrected for motion, off-resonance (susceptibility and eddy-current induced) and slice dropout artefacts using the Eddy tool [9,10]. The DTI model is obtained using the DTIFIT tool within FSL from the b=1000 shell in 50 directions to create measures such as FA and MD. Further, the DTI FA images are then fed into TBSS. Along with the DTI model a NODDI model was also fit to the data, yielding additional parameters such as intra-cellular volume fraction (ICVF, an index of white matter neurite density) and isotropic or free water volume fraction (ISOVF). The two models were used to extract 675 diffusion IDPs over 75 different white matter tract regions. More details on data acquisition and pre-processing can be found in the UKB documentation [17] as well as other studies which address the processing pipelines [18,19].

Normative modeling

In preparation for the modelling stage, the subjects with available dMRI were split into test and training sets. The test set consisted of a random sample of 5000 subjects, while the training set consisted of the remainder of the data (n=22658). The list of diffusion IDPs was selected on the basis of diffusion model (DTI and NODDI) and tract, both categories including simple and complex examples with different modelling difficulties. We chose to model across 4 diffusion parameters. The FA and MD parameters of the DTI model were selected for their relative simplicity since they represent a direct measurement of diffusion influenced by tissue microstructure. On the other hand, we selected the ICVF and ISOVF parameters of the NODDI model for their relative complexity since they enable a more specific characterization of tissue microstructure by estimating neurite density and orientation dispersion. A total of 6 tracts were chosen based on their size and modelling difficulty and comprised: the corpus callosum, the corticospinal tract (both left and right), the uncinate fasciculus (both left and right) and the fornix. This yields a list of 24 IPDs in total.

A normative model was trained with the training set to estimate the normal range of each IDPs value according to age. To account for the possible non-linear effects and non-Gaussian distributions within the dataset, we used a warped Bayesian linear regression (BLR) model [20], as also used in prior work [20,21]. Specifically, this involves applying a third order polynomial B-spline basis expansion over age with 5 evenly spaced knots with SinhArcsinh warping. Next, the test set was used to estimate each subjects' deviation from the normal range of each IDP by computing the individual z-score. The fit statistics of the model were computed including explained variance, skew, and kurtosis. The models were refitted after the exclusion of the outliers from the dataset in order to assess the effect the outliers had on the model fit. The amount of deviation for each subject was visualized by plotting the individual z-scores across the mean and centiles of variation predicted by the model. For a detailed introduction to warped BLR normative modeling please consult the dedicated paper by Fraza et al. [20]. All the statistical analyses were performed in Python version 3.8, using the PCN toolkit [22].

Normative modeling outliers

The outliers of the normative models consist of the participants who presented a z-score larger than 7 standard deviations from the mean in any of the IDPs model. We used a high threshold to select the most severe artefact cases and lower the risk of including participants with anatomical anomalies. The outlier images were visually assessed and labeled according to the artefact type present. For an accurate label the following scans were inspected: b0 scan (to assess the general quality), T1 scan (to assess the anatomical integrity), skeletonized FA map (to assess registration/model estimation), the FA warp from dMRI space to standard space (to assess the registration). We also computed the frequency with which one outlier appears across the IDPs as well as the number of artefact types present in each IDP.

Visual and Quantitative quality control

The visual quality control (VQC) consists of assigning each subject a score from 1 to 3 where 1 = no artefact, 2 = slight artefact, 3 = severe artefact. In order to accelerate the process, 9 slices per subject were used for the visual assessment, 3 from each view (axial, coronal, sagittal) of the B0 images with a T1 masks (in dMRI space) overlay. Two experienced raters were trained to perform the manual labeling on a 500 sample dataset by following a locally developed protocol which established the types and severity of the artefacts and how to provide a score accordingly. The visually captured artefacts included out of field of view scans, signal loss, brain extraction errors and susceptibility distortions. The inter-class correlation (ICC) metric was then used to assess the agreement between raters.

The quantitative quality control measurements consist of 21 quality descriptors for diffusion data including 18 parameters obtained with the Eddy QC tool [10], 2 parameters directly available from UKB and one locally derived parameter. This list includes motion parameters (average and relative motion, translation, rotation), number of outliers per slice (for assessing signal dropout), SNR and angular CNR, T1 vs DWI discrepancy and the discrepancy (i.e., registration cost) between the FA image in MNI space and the FMRIB FA template. A full list of the acquired measurements and their source is presented in Table 1.

The labels obtained with the visual QC were used as ground truth for training a linear SVM classifier with the quantitative QC measurements. We separated the scores in two classes where scores 1 and 2 were grouped together into a single class representing acceptable scans (two classes: 0 = acceptable, 1 = not acceptable). Before classification, the data was balanced by subsampling the artifact-free class in order to prevent biasing the performance results. The data balancing consisted of creating 20 random subsamples of the artefact-free data that match the sample-size of the artifact data (which varies each round). The classifier was run iteratively 20 times and the classification performance (measured by accuracy, sensitivity, and specificity) obtained using 5-fold cross validation was averaged across the 20 iterations to obtain the final results.

In order to assess the image quality of the outliers and determine the severity and the type of artefacts present we applied the same visual and quantitative QC protocol to the normative modeling outliers. Furthermore, we also identified the quantitative QC outliers (threshold of 2 standard deviations from the mean) and again computed the number of times each outlier appears across all objective QC parameters. A correlation matrix was computed using this information which consisted of the following measurements: T1 discrepancy values, FA vs DWI discrepancy, SNR, Normative Modeling (NM) outlier frequency, AQC outlier frequency and VQC score in order to establish correlations between quality descriptor, visual evaluation and the involved IDP.

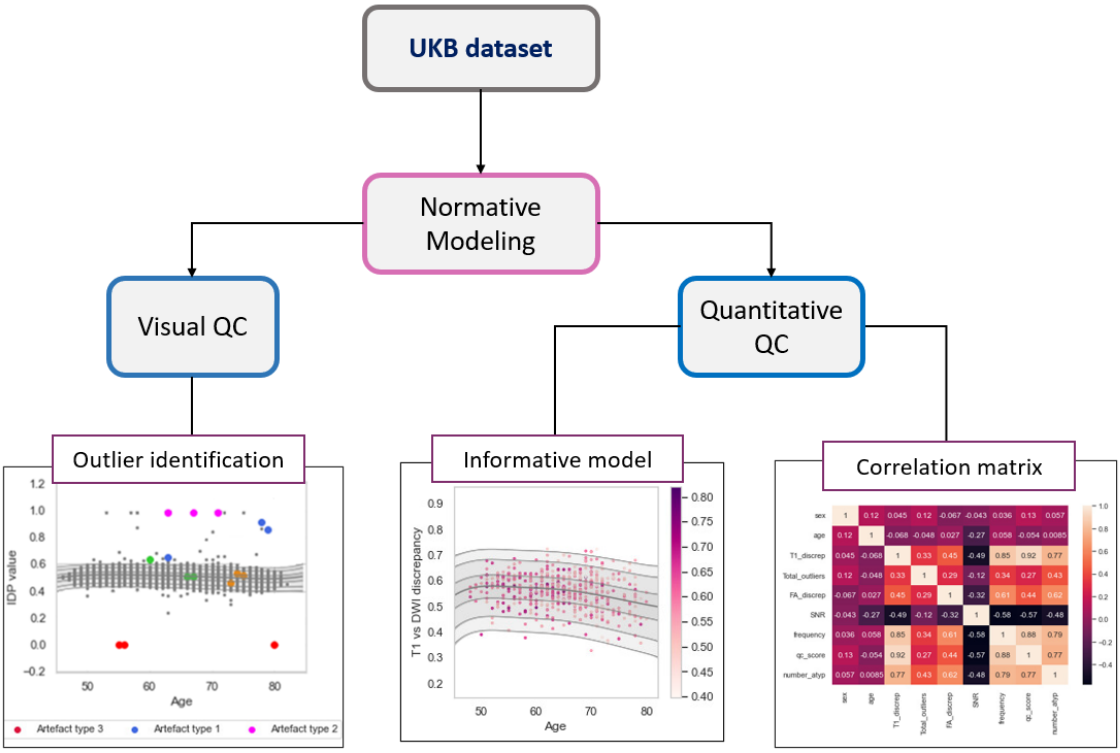


Figure 1. Flowchart describing the main methods used in this paper. 24 diffusion IDPs provided by UK Biobank are used to fit normative models. The estimates provided by the models are used in conjunction with the visual QC protocol to identify and label artefact presence in the scans. Quantitative measures describing the quality of the diffusion data are also used together with the normative model estimates to in order to establish correlations between objective QC parameters and the outlier presence

Table1: List of quantitative dMRI QC metrics obtained using automated QC methods

QC metric	**Description**	**Source**
Average absolute motion (mm)	Volume-to-volume motion	EDDY QC
Average relative motion (mm)	Volume-to-volume motion	EDDY QC
Average x translation (mm)	Volume-to-volume motion	EDDY QC
Average y translation (mm)	Volume-to-volume motion	EDDY QC
Average z translation (mm)	Volume-to-volume motion	EDDY QC
Average x rotation (deg)	Volume-to-volume motion	EDDY QC
Average y rotation (deg)	Volume-to-volume motion	EDDY QC
Average z rotation (deg)	Volume-to-volume motion	EDDY QC
Average stdev x translation (mm)	Within volume motion	EDDY QC
Average stdev y translation (mm)	Within volume motion	EDDY QC
Average stdev z translation (mm)	Within volume motion	EDDY QC
Average stdev x rotation (deg)	Within volume motion	EDDY QC
Average stdev y rotation (deg)	Within volume motion	EDDY QC
Average stdev z rotation (deg)	Within volume motion	EDDY QC
Stdev x axis linear term (Hz/mm)	Eddy current-induced distortion	EDDY QC
Stdev y axis linear term (Hz/mm)	Eddy current-induced distortion	EDDY QC
Stdev z axis linear term (Hz/mm)	Eddy current-induced distortion	EDDY QC
Total number of outliers (%)	Slice drop-out	EDDY QC
Volume-wise number of outliers (%)	Slice drop-out	EDDY QC
Outliers per b-shell (%)	Slice drop-out	EDDY QC
Outliers per PE dir (%)	Slice drop-out	EDDY QC
Average b0 SNR	General scan quality	EDDY QC
Average diffusion CNR per shell	General scan quality	EDDY QC
Number of slices corrected	Slice drop-out	UK Biobank
T1 vs dMRI discrepancy	Geometric distortions	UK Biobank
FA discrepancy	Geometric distortions	Locally derived

Results

Normative modeling

The normative modeling fit was measured in terms of explained variance (EV), skew and kurtosis. Across the 24 IDPs the models had mean EV of $\mu_{EV} = 0.092$ with standard deviation $\sigma_{EV} = 0.089$, mean Skew of $\mu_{Skew} = 0.433$ with standard deviation $\sigma_{Skew} = 0.686$, mean Kurtosis of $\mu_{Kurtosis} = 1.482$ with standard deviation $\sigma_{Kurtosis} = 1.886$. The fit of the models after the exclusion of the outliers was not statistically significant ($p = 0.33$).

The normative models were fit to the 24 diffusion IDPs and the subjects with a z-score beyond 7 standard deviations away from the mean were considered outliers. Across all IDPs there were a total of 85 unique outlier subjects, out of which 64 appeared in at least two IDPs. After visually inspecting each outlier, we devised 3 main categories of image errors:

1. The first category consists of acquisition artefacts: out of the field of view (out of FOV), blur bias field-induced signal inhomogeneities, motion-induced signal dropout etc. These errors are generally easy to spot using a visual QC protocol as there are relatively clear indicators of artefact presence.
2. The second category is pre-processing errors which mainly comprise of registration or brain extraction errors which can occur for various reasons. Importantly, these errors often arise even from good quality data, e.g., in the case of misregistration due to large ventricles, severe atrophy or anatomical anomalies. Therefore, visual QC earlier in the processing pipeline is generally not efficient in identifying these processing errors. In this analysis, further steps were taken to identify these errors involving overlaying the skeletonized FA scan on the Johns Hopkins University (JHU) atlas to check for tract misalignment and visual inspection of the FA warp to MNI space to detect any abnormal distortions pointing to a faulty registration.
3. The third category contains incidental findings which may be present in the dataset. Although this is not an artefact or error, it is important to identify anatomical anomalies and have the possibility to review such participants to decide whether they must be excluded or not from further analyses.

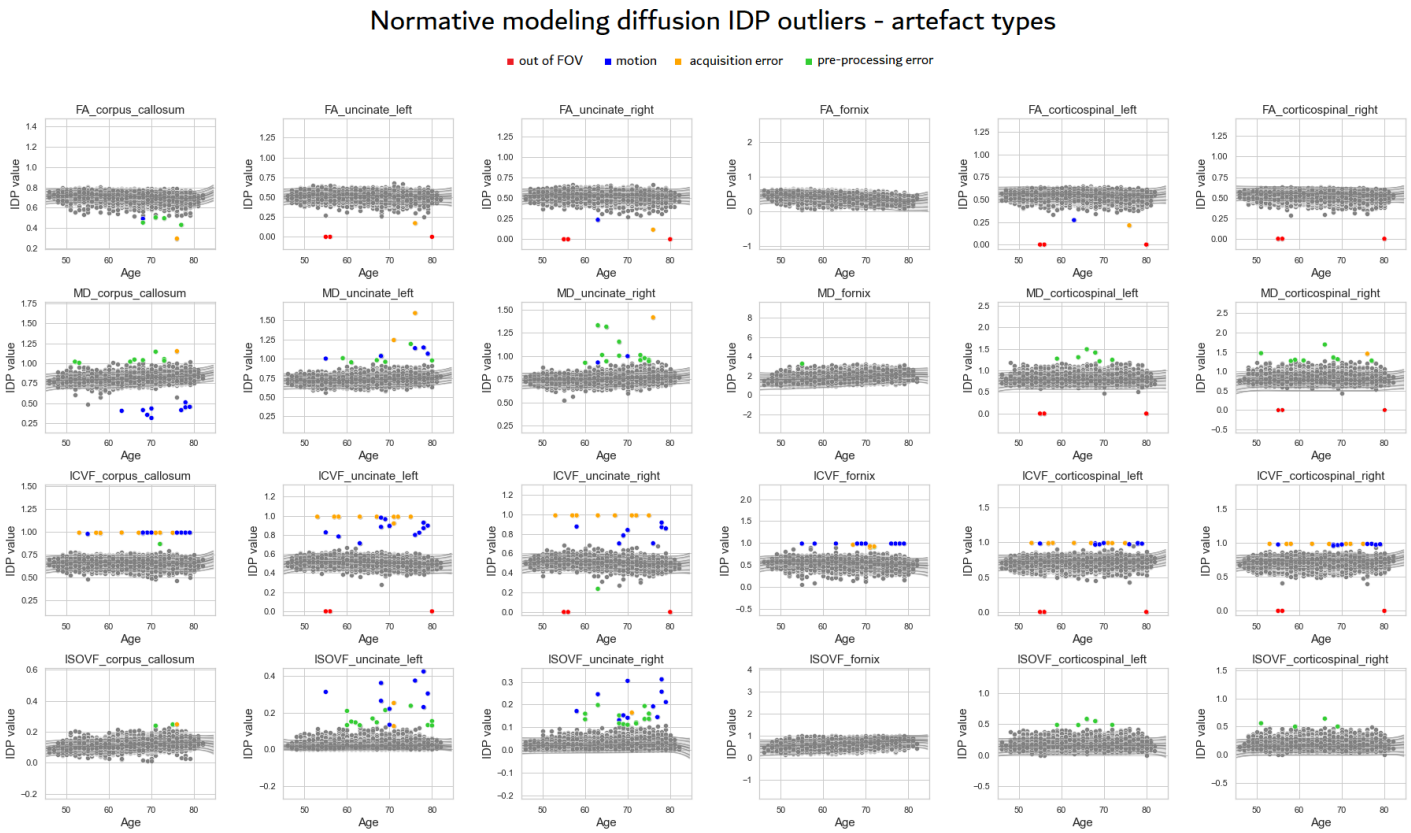


Figure 2. Normative model centiles of 24 diffusion IDPs we trained with the normative models as well as the outliers for each model. The outliers are color coded by artefact type, to highlight their position and frequency within each IDP centile plot. The model outliers cluster according to artefact type while also indicating that artefact presence varies depends on DWI parameter and tract.

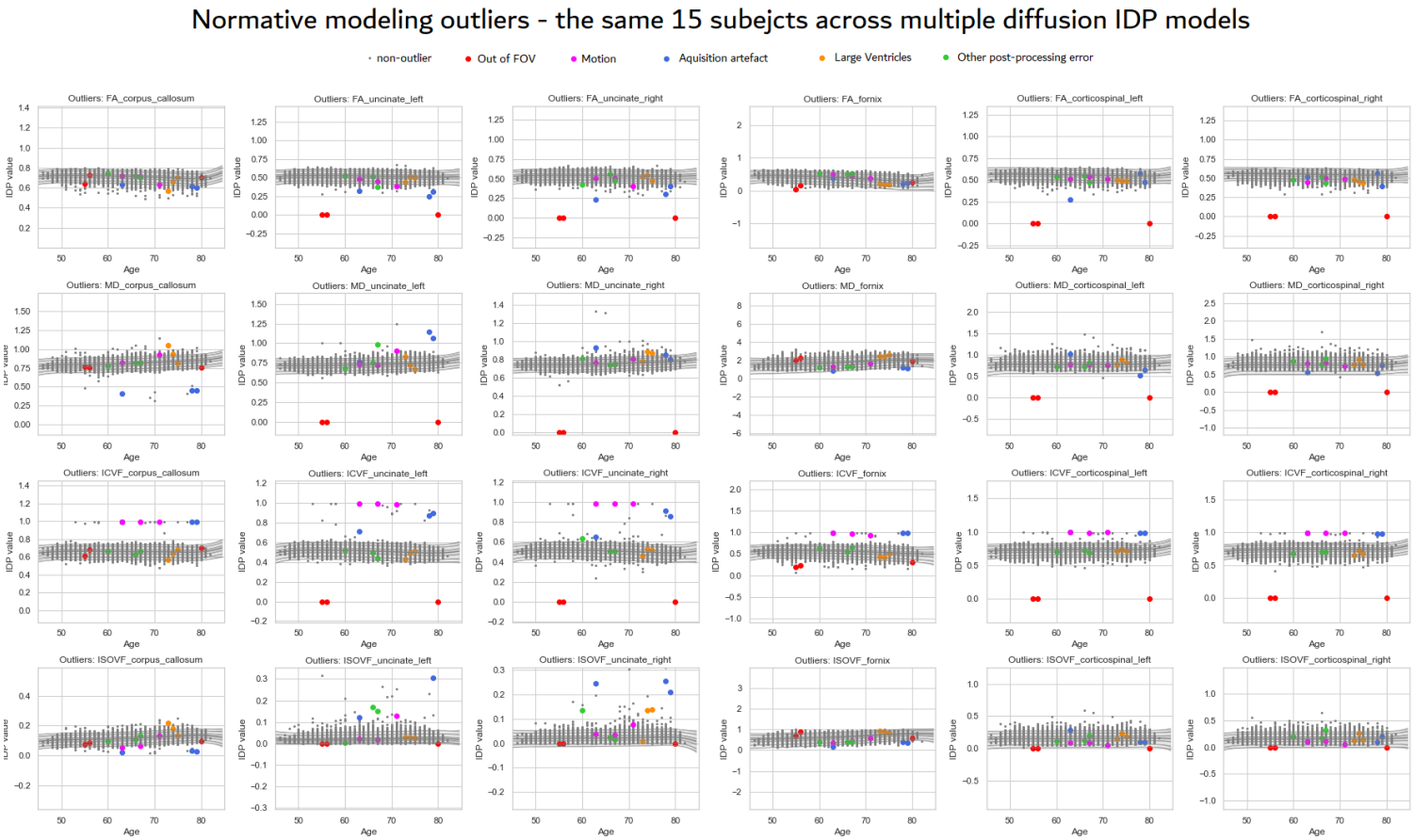


Figure 3. Normative model centiles of 24 diffusion IDPs. The same 15 outliers are color coded according to artefact type. Although the model outliers form cluster patterns according to artefact type, it is visible that while some subjects were identified by the model as extreme outliers in some IDPs, they remain close to the mean in other IDPs. This image corroborates the fact that some IDPs are much more robust than others in terms of tolerating image artefacts and/or preprocessing errors.

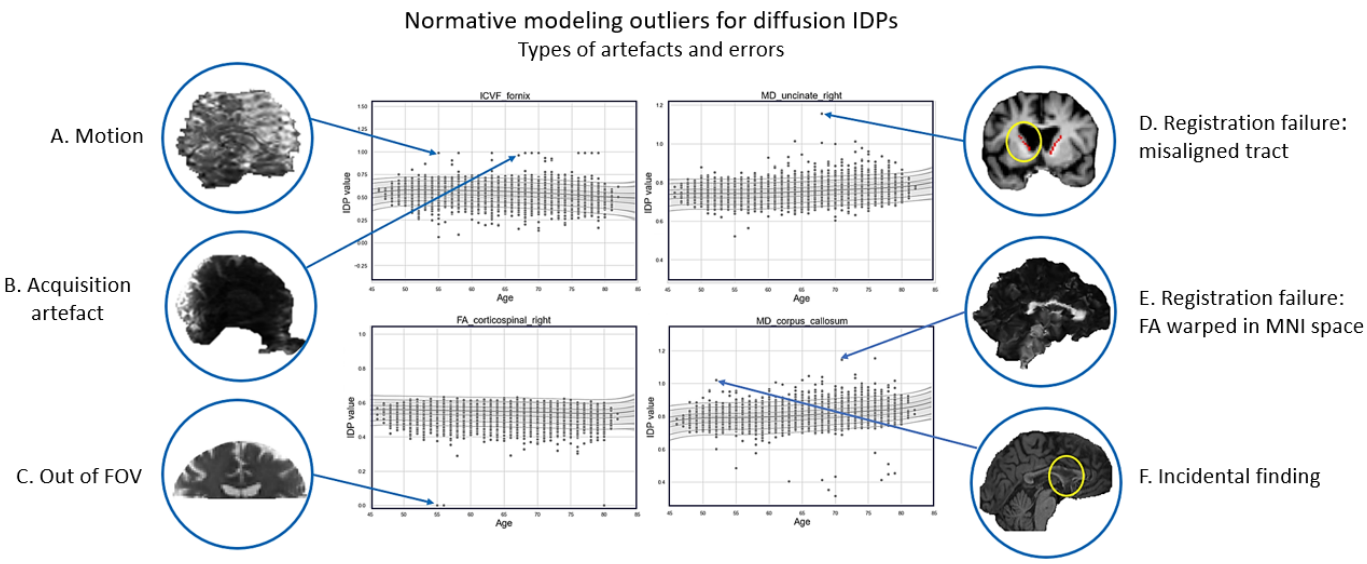


Figure 4. Example of each artefact type up close. On the left side 4 A, 4 B and 4 C are examples of image artefacts which can easily be spotted during visual QC. On the right side, 4 D and 4 E presents examples of scans with pre-processing errors. 4 F is an example of an incidental finding which was detected by the normative model.

Figure 4 D is an example of registration failure in a subject with enlarged ventricles who appears as an outlier in in the 'Mean MD in uncinate fasciculus right' IDP. This example is also illustrated in Figure 5, where the b0 image as well as the FA image is also included. The error can be seen in Figure 5C from the overlay of the skeletonized uncinate fasciculus tract within the JHU atlas (in DWI native space) on the T1 scan of the subject (in DWI native space). It is noticeable from the image that the right uncinate fasciculus is positioned in gray matter. This indicates a faulty registration and explains why this subject has an abnormal value within the 'Mean MD in uncinate fasciculus right' IDP.

Another example of faulty registration is shown in Figure 4 E which consists of the FA scan warped in MNI space of a subject who appears to be an outlier in the 'mean MD in corpus callosum' IDP. It is apparent from the image that the registration is extremely distorted which has a big impact on the appearance of the corpus callosum, thus explaining the abnormal value of this IDP in this subject. It is important to note that these two examples could not be identified as artefactual by visually inspecting any of the DWI scans from the visual protocol. The last example in Figure 4 F shows the T1 scan of a subject with an abnormal corpus callosum which has been identified as an outlier in the 'mean MD in corpus callosum'. This anatomical abnormality is an incidental finding.

Outlier MD uncinate fasciculus (right): misaligned tract

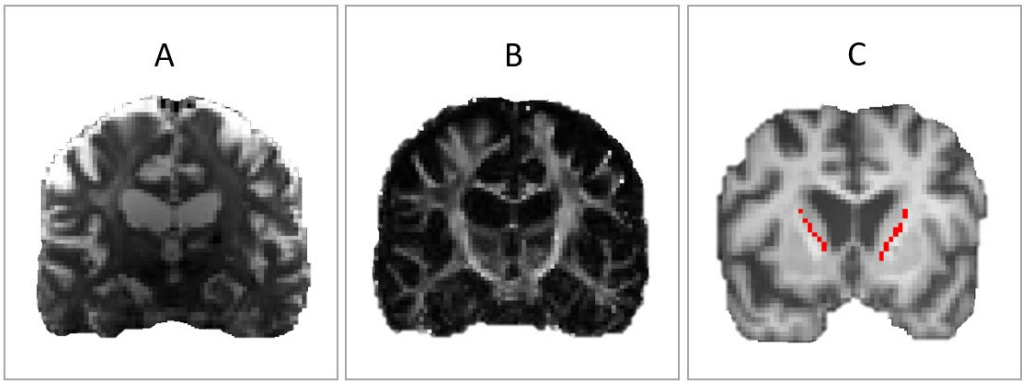


Figure 5. Axial slice of the same subject represented by three different data types. 5A shows the b0 image. 5B shows the FA image while 5C shows the skeletonized uncinate fasciculus (red color) overlayed on the T1 image. Although 5A and 5B have a normal appearance and there is no indication of artefact presence, a closer inspection of the registration reveals that the uncinate fasciculus (right) tract is misaligned. This suggest that these types of errors are not detectable by regular visual QC.

Visual and Quantitative quality control

Visual QC was performed on a sample of 500 subjects by two raters who gave scores according to the artefact severity from 1 to 3. The raters had an interclass correlation coefficient of 0.71 which indicates a moderate reliability. In Figure 6.A , SNR values were color coded in the normative model centile plot of one of the IDPs (mean MD in uncinate fasciculus left). There appears to be little correlation between the quantitative QC outliers and the normative modeling outliers. The visual QC scores were used in several analyses to assess the correlation between visual and quantitative QC measures. In Figure 6B, the visual QC scores of rater 1 were color coded in the normative model centile plot of one of the IDPs (mean MD in uncinate fasciculus left). There also appears to be little correlation between the bad visual QC scans (score = 3) and the normative modeling outliers (data points further away from the mean). Figure 6C also shows a plot of the pairwise relationships between several quantitative QC parameters (T1 discrepancy, SNR and CNR b=1000) and the visual QC scores which are color coded in the scatterplots. The distributions of the scores overlap consistently and the scatterplots do not show any clear clustering of the data points by score. This suggests that the correlation between the visual and automated QC is weak. Figure 6 D shows the performance scores of the SVM classifier which was used to distinguish between acceptable (score 1 and 2) ad unacceptable scans (score 3). The accuracy scores are no higher than 0.6 in both rates which corroborates that the link between visual and automated QC is not significant.

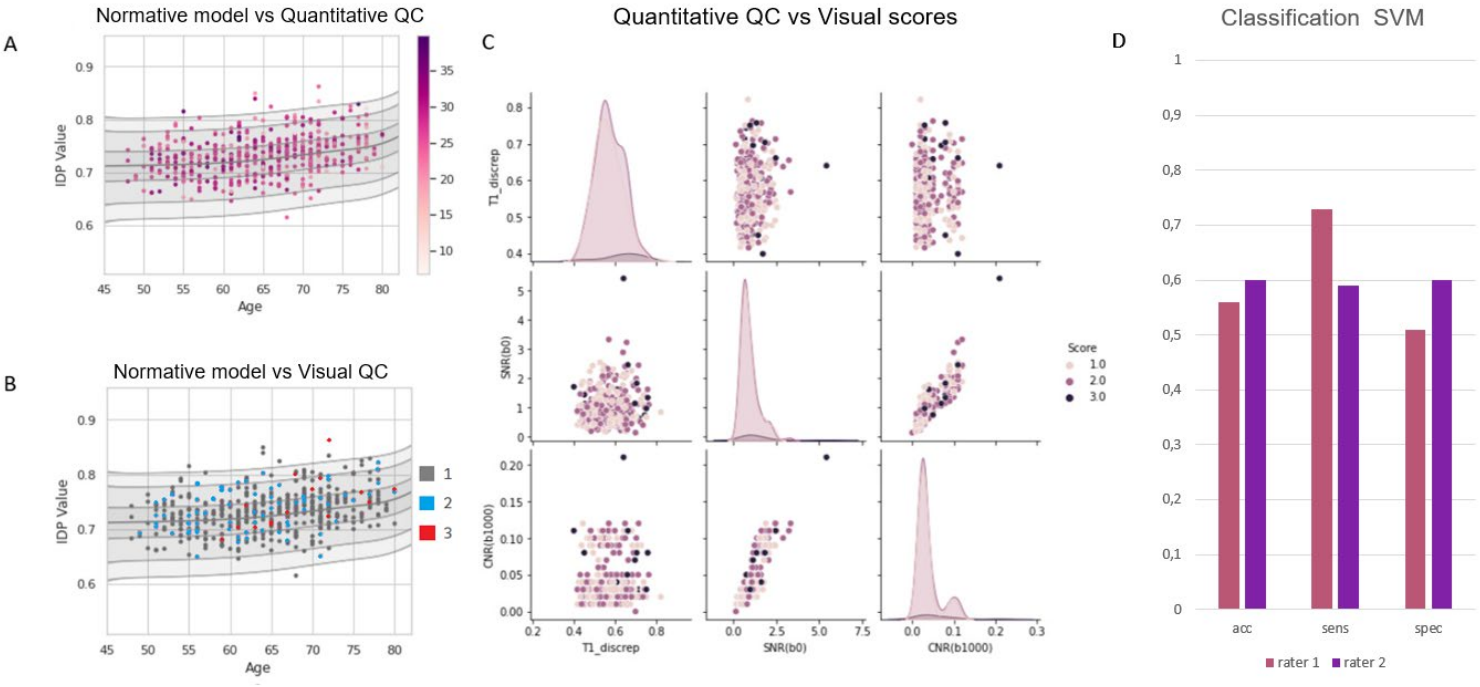
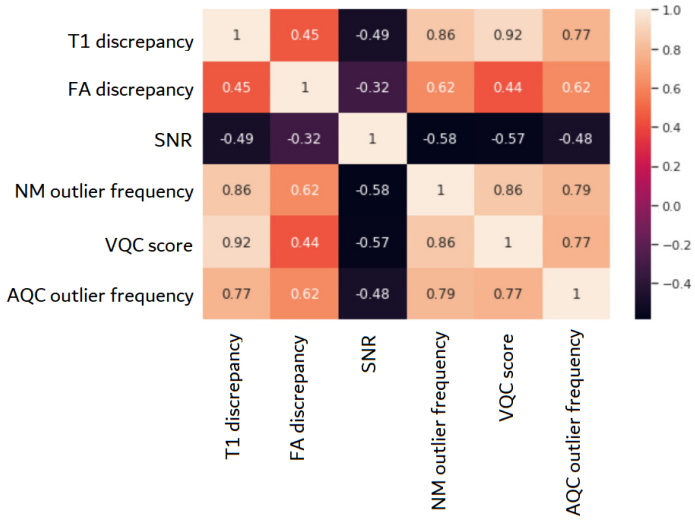


Figure 6 A. SNR values of the 500 subjects color coded into the normative model centile plot of the 'mean MD in uncinate fasciculus left' IDP. B. Visual QC scores of the 500 subjects color coded into the normative model centile plot of the 'mean MD in uncinate fasciculus left' IDP. C. Pairplot between three quantitative QC parameters and the visual QC scores showing the pairwise relationship between visual and automate QC modalities. D. SVM classification performance scores for both raters.

Using the 85 outlier subjects, a correlation matrix was computed with select measurements obtained from both the visual QC and quantitative QC parameters as well as parameters such as frequency with which a subject appears as outlier in the 24 IDP normative models. This can be seen in Figure 7. In this case there is a strong positive correlation between the visual QC scores and the T1 discrepancy

parameter as well as the outlier frequency in both the normative model and the quantitative QC. This suggests that, in the case of extreme outliers, the correlation between QC modalities is much stronger.



Correlation matrix between Automated QC (AQC) metrics (T1 discrepancy, FA discrepancy, SNR and AQC outlier frequency), normative modeling (NM) outlier frequency and Visual QC (VQC) score. This correlation was computed using the 85 outliers detected by the normative model and shows the relationship between the three QC methods in case of images heavily affected by artefacts and postprocessing errors.

Discussion

In this paper we propose a new semi-automated pipeline for dMRI quality control. By training a normative model with 24 diffusion IDPs and analyzing the outliers, we were able to identify severe artefact and pre-processing errors which were not detected by other QC methods such a visual QC protocol or an automated QC tool like EDDY QC. Therefore, the time and effort invested in assessing the quality of a large dMRI dataset was significantly reduced. Moreover, that we are able to detect preprocessing errors that may occur even in the absence of artefacts and that may go undetected in conventional QC workflows. At the same time, we showed the wide discrepancy between QC modalities and demonstrated that visual QC is subjective and prone to failure since artefacts can be subtle or at time impossible to detect visually.

It is important to note that artefact detection is dependent on both the type of derived measure and the tract from which it is derived (i.e. regional specificity). In Figure 2, we see that the FA metric is robust, showing very little outliers. On the other hand, all of the ICVF metric is very sensitive to image errors, showing many outliers and a high amount of artefact types. Generally, larger tracts such as the corpus callosum and the corticospinal tract are easier to model and have relatively fewer outliers than smaller tracts. The only exception is the fornix, which –perhaps surprisingly– has almost no outliers (except the ICVF), this is because, when looking at the measurement range of the fornix tract IDPs, we can point out the very large variance. This means that the fornix, being a small and hard to model tract is often misregistered which makes this phenomenon not as evident as in the others. When looking at artefact types, it is worth pointing out that MD and ISOVF appear to be more sensitive to pre-processing errors.

When following the same 15 subjects across all IDP models, we see that in many cases the subjects which are outliers in some IDPs are perfectly normal, and quite close to the mean in other IDPs. In Figure 3, we color coded the subjects by artefact type. On the other hand, pre-processing errors (either with enlarged ventricles or not) are less evident and can sometimes only be picked up by the affected tract. This is shown in Figure 3, where an outlier in the 'mean MD in uncinate right' shows a visible registration failure in the uncinate fasciculus (right) tract. This suggests that, although image errors may be present in many cases, it is not necessary to remove all faulty scans, since sometimes, it may not make a difference in the analysis. This is true in case of FA in corpus callosum, which is a very robust IDP and does not require extensive clean-up. However, in instances such as the ICVF of the uncinate fasciculus, a serious quality assessment may be required.

The normative modeling approach to quality control has the advantage of being able to detect pre-processing errors, as we have shown in Figure 2, 3 and 4, and can be applied easily to any IDP of interest. This is an important achievement due to the high impact these errors can have on the results of an analysis on the affected tract but also due to the possibility to correct many of these errors. Conversely, the scans which contain severe image artefacts can often not be corrected and must be discarded. We have seen in the Literature review section of this paper, that there are several toolboxes dedicated to detecting and correcting artefacts as part of the dMRI pre-processing pipeline (e.g., FLS Eddy [9]). However, these algorithms have limitations and a testament to these limitations is the relatively large number of artefacts that are still detected in the UK Biobank dataset, which has already been pre-processed and corrected for distortions. Unlike the image artefacts, the pre-processing errors may be resolvable by adjusting the processing pipelines potentially allowing the data to still be used.

The visual and automated QC was carried out for a sample of 500 UK Biobank subjects in order to assess the compatibility of the methods. Specifically, we wanted to see if the same outliers occur for all methods and evaluate the amount of overlap. Figure 6 A, B and C show that there is very little (visual) overlap between the outliers of the three methods (visual, quantitative, and normative modeling). The performance of the SVM classifier for both raters also suggests a weak link between the manual labels and quantitative parameters. However, the correlation matrix in Figure 7, which was computed on the 85 outliers, tells us that in extreme cases, the three modalities have a close relationship and can explain each other's variance. We would like to suggest that these results do not necessarily mean that the three methods are incompatible in the case of less severe artefacts but that that are actually complementary. A complete and detailed quality assessment of diffusion MRI scans (if necessary) would make use of all these methods which are designed to detect different artefact types. It would be fair to say, however, that the normative modeling approach to QC is a comprehensive method which can provide both qualitative and quantitative results depending on how it is used and the scope of the study.

A strength of the normative modeling approach is that it does not need any manually labeled data. Many QC methods which are based on classification require accurate ground truth data which must be acquired through a visual QC protocol. We have shown in this study that inter-rater reliability is not perfect, which although seems to be a limitation, only reflects the reality of manual labeling – it is imperfect and subjective. Therefore, carrying out a visual QC protocol is undesirable and makes the classification method unlikely to be preferred. Classification based methods also have the shortcoming of being unable to identify artefacts which are not present in the training set. This is not the case for the normative modeling method, which is able to detect unseen errors. Furthermore, our approach does not suffer from the shortcomings of unbalanced datasets which present a problem in any method which uses artefact vs artefact-free classes. The nature of the dMRI dataset (and most datasets in the medical imaging field) is that the majority of the scans will be artefact-free, creating a very large imbalance between classes which must be carefully handled to overcome bias in classification performance scores. The normative modeling approach is able to assess the level of abnormality at the individual level rendering categorical class labels unnecessary. This is a powerful point since class division puts limitations on interindividual variability and suppresses the dimensionality of the data.

The main limitation of this study is the lack of an existing comparable method for validation purposes. The best QC modality is subjective and depends on the analysis that will be conducted. For example, detecting group differences via a univariate group probably has a lower sensitivity to artefacts than a biomarker detection task using a classifier. The perfect QC depends on the goal of the study and the type of the data as well as on the size of the dataset and the time available for clean-up. Other shortcomings regard the limited number of chosen IDPs. The QC of the dataset could have been more detailed and accurate if more IDPs were used in the analysis and if the outlier threshold was lower, allowing us to catch the less severe artefacts. However, for the purposes of showing the efficacy of the method, we had to limit ourselves to the essential number of IDPs and a high threshold to detect the most obvious outliers.

In conclusion, this paper introduces a new method for dMRI quality control. The normative modeling approach to QC is a semi-automated pipeline which is able to detect subjects who present image artifacts as well as other pre-processing errors. The analysis is performed at an individual level, overcoming the shortcomings of class division as well as having the ability to detect new, unseen artefacts while not requiring any manually labeled data. Moreover, in this study we showed, with the help of our method, that there are two main categories of image errors: image artifacts which can be detected visually but are irreversible and pre-processing errors which usually go undetected but can

potentially be fixed. The normative modeling approach can be used together with other QC methods such as a visual QC protocol or other objective QC parameters and can also be customized for detection severity (by changing the outlier threshold) based on the scope of the study. Finally, we showed that although other QC methods are not consistent between each other nor with the normative modeling method, they can be complementary and enhance the efficacy of the overall quality assessment.

References

1. Le Bihan D. Looking into the functional architecture of the brain with diffusion MRI. Nature reviews neuroscience. 2003 Jun 1;4(6):469-80.
2. Le Bihan, D., Poupon, C., Amadon, A., & Lethimonnier, F. (2006). Artifacts and pitfalls in diffusion MRI. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *24*(3), 478-488.
3. Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., & Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *13*(4), 534-546.
4. Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC. NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage. 2012 Jul 16;61(4):1000-16.
5. Rosen, A.F.G., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., Ciric, R., Cook, P.A., Davatzikos, C. and Elliott, M.A., 2018. Garcia de 423 La Garza A, Gennatas ED, Quarmley M, Schmitt JE, Shinohara RT, Tisdall MD, Craddock RC, Gur RE, Gur RC, 424 Satterthwaite TD. Quantitative assessment of structural image quality. *NeuroImage*, *169*, pp.407-418.
6. Chang LC, Walker L, Pierpaoli C. 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. 2012 Nov;68(5):1654-63.
7. Zhang B, Li S, Feng G, Qian Z, Zhang X. Patch Diffusion: a general module for face manipulation detection. InProceedings of the AAAI Conference on Artificial Intelligence 2022 Jun 28 (Vol. 36, No. 3, pp. 3243-3251).
8. Oguz I, Farzinfar M, Matsui J, Budin F, Liu Z, Gerig G, Johnson HJ, Styner M. DTIPrep: quality control of diffusion-weighted images. Frontiers in neuroinformatics. 2014 Jan 30;8:4.
9. Andersson JL, Sotiropoulos SN. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage. 2016 Jan 15;125:1063-78.
10. Andersson JL, Graham MS, Zsoldos E, Sotiropoulos SN. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. Neuroimage. 2016 Nov 1;141:556-72.
11. Bastiani M, Cottaar M, Fitzgibbon SP, Suri S, Alfaro-Almagro F, Sotiropoulos SN, Jbabdi S, Andersson JL. Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction. Neuroimage. 2019 Jan 1;184:801-12.
12. Samani ZR, Alappatt JA, Parker D, Ismail AA, Verma R. QC-Automator: Deep learning-based automated quality control for diffusion mr images. Frontiers in neuroscience. 2020 Jan 22;13:1456.
13. Ahmad A, Parker D, Dheer S, Samani ZR, Verma R. 3D-QCNet–A pipeline for automated artifact detection in diffusion MRI images. Computerized Medical Imaging and Graphics. 2023 Jan 1;103:102151.
14. Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biological psychiatry. 2016 Oct 1;80(7):552-61.
15. Ollier W, Sprosen T, Peakman T (2005). UK Biobank: from concept to reality. Pharmacogenomics 6: 639–646

16. UK Biobank (2006). Protocol for a large-scale prospective epidemiological resource. www.ukbiobank.ac.uk/resources/.
17. Smith S.M., Alfaro-Almagro F. and Miller K.L. UK Biobank Brain Imaging Documentation version 1.8.
18. Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JL, Griffanti L, Douaud G, Sotiropoulos SN, Jbabdi S, Hernandez-Fernandez M, Vallee E, Vidaurre D. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage. 2018 Feb 1;166:400-24.
19. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JL, Griffanti L. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nature neuroscience. 2016 Nov;19(11):1523-36
20. Fraza CJ, Dinga R, Beckmann CF, Marquand AF. Warped Bayesian linear regression for normative modelling of big data. Neuroimage. 2021 Dec 15;245:118715.
21. Rutherford S, Fraza C, Dinga R, Kia SM, Wolfers T, Zabihi M, Berthet P, Worker A, Verdi S, Andrews D, Han LK. Charting brain growth and aging at high spatial precision. Elife. 2022 Feb 1;11:e72904.
22. https://github.com/amarquand/PCNtoolkit