Fast qualitY conTrol meThod foR derIved diffUsion Metrics (YTTRIUM) in big data analysis: U.K. Biobank 18,608 example

Ivan I. Maximov, Dennis van der Meer, Ann-Marie G. de Lange, Tobias Kaufmann, Alexey Shadrin, Oleksandr Frei, Thomas Wolfers, Lars T. Westlye

1Department of Psychology, University of Oslo, Oslo, Norway
2Norwegian Centre for Mental Disorders Research (NORMENT), Department of Mental Health and Addiction, Oslo University Hospital and Institute of Clinical Medicine, University of Oslo, Oslo, Norway
3Department of Health and Functioning, Western Norway University of Applied Sciences, Bergen, Norway
4School of Mental Health and Neuroscience, Faculty of Health Medicine and Life Sciences, Maastricht University, The Netherlands
5LREN, Centre for Research in Neurosciences – Department of Clinical Neurosciences, CHUV and University of Lausanne, Lausanne, Switzerland
6Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford, UK
7Center for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway
8KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway

Correspondence
Ivan I. Maximov, NORMENT, Department of Psychology, University of Oslo, Oslo, Norway. Email: ivan.maximov@psykologi.uio.no

Funding information
Norges Forskningsråd, Grant/Award Number: 249795

Abstract
Deriving reliable information about the structural and functional architecture of the brain in vivo is critical for the clinical and basic neurosciences. In the new era of large population-based datasets, when multiple brain imaging modalities and contrasts are combined in order to reveal latent brain structural patterns and associations with genetic, demographic and clinical information, automated and stringent quality control (QC) procedures are important. Diffusion magnetic resonance imaging (dMRI) is a fertile imaging technique for probing and visualising brain tissue microstructure in vivo, and has been included in most standard imaging protocols in large-scale studies. Due to its sensitivity to subject motion and technical artefacts, automated QC procedures prior to scalar diffusion metrics estimation are required in order to minimise the influence of noise and artefacts. However, the QC procedures performed on raw diffusion data cannot guarantee an absence of distorted maps among the derived diffusion metrics. Thus, robust and efficient QC methods for diffusion scalar metrics are needed. Here, we introduce Fast qualitY conTrol meThod foR derived diffUsion Metrics (YTTRIUM), a computationally efficient QC method utilising structural similarity to evaluate diffusion map quality and mean diffusion metrics. As an example, we applied YTTRIUM in the context of tract-based spatial statistics to assess associations between age and kurtosis imaging and white matter tract integrity maps in U.K. Biobank data (n = 18,608). To assess the influence of outliers on results obtained using machine learning (ML) approaches, we tested the effects of applying YTTRIUM on brain age prediction. We demonstrated that the proposed QC pipeline represents an efficient approach for identifying poor quality datasets and artefacts and increases the accuracy of ML based brain age prediction.

KEYWORDS
brain maturation, diffusion QC, DKI, DTI, U.K. Biobank, WMTI, YTTRIUM
1 | INTRODUCTION

Diffusion magnetic resonance imaging (dMRI) provides a range of structural brain features based on routine clinical measurements, which has contributed to its popularity across fields and applications (de Lange et al., 2020; Kochunov et al., 2015; Westlye et al., 2010). Advanced dMRI is technically challenging and often involves time-consuming acquisitions placing high demands on the performance and stability of the scanner hardware. Therefore, dMRI data are vulnerable to experimental setup perturbations including post-processing approaches, which might bias the results. In turn, optimised post-processing pipelines (Ades-Aron et al., 2018; Maximov, Ainaes, & Westlye, 2019; Tournier et al., 2019) and stringent procedures for quality control (QC; Alfaro-Almagro et al., 2018; Bastiani et al., 2019; Graham, Drobnjak, & Zhang, 2018; Haddad et al., 2019) are important to increase reliability and sensitivity. Various approaches have been developed to detect and correct artefacts in raw diffusion data originating, for example, from eddy currents, bulk head motions, susceptibility distortions (Andersson & Sotiropoulos, 2016), noise (Kochunov et al., 2018), Gibbs ringing artefacts (Perrone et al., 2016; Veralta, Fieremans, Jelescu, Knoll, & Novikov, 2016; Veraart, Novikov, et al., 2016), presence of outliers (Koch, Zhukov, Stöcker, Groeschel, & Schultz, 2019) and diffusion metric variability (David, Mesri, Viergever, & Leemans, 2019; Maximov et al., 2015).

However, QC and data harmonisation procedures applied on raw diffusion data (Fortin et al., 2017; Mirzaalian et al., 2018) do not guarantee accurate numerical computation of scalar diffusion metrics. Derived diffusion metrics from diffusion or kurtosis tensors are sensitive to a range of subject-specific factors such as age or various brain disorders, but also to applied numerical algorithm or its programming implementation (David et al., 2019; Grinberg et al., 2017; Lebel et al., 2012; Maximov et al., 2015). The effects of noisy observations on subsequent between-subjects analysis involving the derived diffusion metrics can be mitigated using simple outlier detection procedures (see, e.g., de Lange et al., 2020; Richard et al., 2018; Tannenes et al., 2018). However, few publications have directly assessed the effects of QC filtration of final data and performing a sanity check of the derived scalar maps before the statistical analysis. As an example, one can use a visual inspection (see, e.g., slicesdir utility from FSL [Smith et al., 2007]) or truncation based on variability of the data and their SD. We know that outliers might affect the results of analysis, in particular, machine learning (ML) algorithms and related prediction or classification output. One example is brain age prediction using neuroimaging data (Kaufmann et al., 2019; Smith, Vidaurre, Alfaro-Almagro, Nichols, & Miller, 2019), where corrupted data either in the training or test sets will influence the accuracy of the prediction.

Here, we introduce a QC method for the derived diffusion maps based on twofold parameterisation: first, diffusion data reduction based on the scalar diffusion values averaged across skeleton voxels using tract-bases spatial statistics (TBSS; Smith et al., 2007), and, second, structural similarity (SSIM; Wang, Bovik, Sheikh, & Simoncelli, 2004) of individual diffusion maps relative to the mean diffusion image derived from all subjects. We demonstrate feasibility of this approach for U.K. Biobank (UKB) data (Miller et al., 2016) using three commonly applied diffusion approaches: diffusion tensor imaging (DTI) (Basser, Mattiello, & Lebihan, 1994), diffusion kurtosis imaging (DKI) (Jensen, Helpern, Ramani, Lu, & Kaczenski, 2005) and white matter tract integrity (WMTI; Fieremans, Jensen, & Helpern, 2011). We evaluated the effect of the developed QC approach by assessing age-diffusion associations and the accuracy of brain age prediction using ML technique.

2 | METHODS AND MATERIALS

2.1 | Participants and MRI data

We used dMRI data obtained from 18,608 subjects (see Figure 1 for age and sex distribution). An accurate overview of the UKB imaging acquisition parameters and initial QC pipeline can be found in Alfaro-Almagro et al. (2018) and Miller et al. (2016). Briefly, a conventional Stejskal-Tanner monopolar spin-echo echo-planar imaging (EPI) sequence was used with multiband factor 3, diffusion weightings (b-values) were 1 and 2 ms/μm² and 50 non-coplanar diffusion directions per shell. All subjects were scanned at 3T Siemens Skyra scanners with a standard Siemens 32-channel head coil, in Cheadle and Newcastle, UK. The spatial resolution was 2 mm³ isotropic, and five AP versus three PA images with b = 0 ms/μm² were acquired. All diffusion data were post-processed using an optimised diffusion pipeline (Maximov et al., 2019) consisting of six steps: noise correction (Veraart, Fieremans, et al., 2016; Veraart, Novikov, et al., 2016), Gibbs-ringing correction (Kellner, Dhital, Kiselev, & Reisert, 2016), estimation of echo-planar imaging distortions, head motions, eddy-current and susceptibility distortions (Andersson & Sotiropoulos, 2016), spatial smoothing using fslmaths from FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) with a 1 mm³ Gaussian kernel, and diffusion metrics estimation using Matlab scripts (MathWorks, Natick, MA; Veraart, Sijbers, Sunaert, Leemans, & Jeurissen, 2013). UKB data were processed using the high-performance computing facility Colossus.
at the University of Oslo and large data storage located at Services for
Sensitive Data (TSD).

2.2 | Diffusion metrics

2.2.1 | DTI and DKI

Diffusion signal decay can be represented as the Taylor expansion
along diffusion weightings (Novikov, Kiselev, & Jespersen, 2018). This
can be approximated by two diffusion tensors of the second (DTI) and
fourth (DKI) orders of diffusion wavevector. A set of scalar maps are
derived from eigenvalues of the both tensors such as FA, mean, axial
and radial diffusivities (MD, AD, RD, respectively), and mean, axial and
radial kurtosis (MK, AK and RK, respectively). The scalar maps charac-
terise integrative features of brain tissue with potential to represent
sensitive biomarkers (Jones, 2010).

2.2.2 | WMTI

In the frame of standard diffusion model (Novikov, Kiselev, &
Jespersen, 2018), WMTI represents an intra-axonal space as a bundle
of cylinders with effective radius equals to zero (Fieremans et al., 2011). The cylinders are impermeable, that is, there is no water
exchange between intra- and extra-axonal spaces. The extra-axonal
space is described by anisotropic Gaussian diffusion. In order to keep
the model simple a few more assumptions have been made: intra-
axonal space consists of mostly myelinated axons without any
contribution from myelin due to a fast relaxation rate across of typical
diffusion times; at the same time in extra-axonal space the glial cells
possess fast water exchange with extra-cellular matrix; both intra-
and extra-axonal spaces are modelled by Gaussian diffusion tensors.
In order to avoid degeneration (Jelescu, Veraart, Fieremans, & Novikov,
2016), intra-axonal diffusion is assumed to be slower than diffusion in
extra-axonal matrix. However, this assumption should be considered
carefully, because it artificially reduces a set of plausible estimations
appearing in the conventional diffusion experiments (Novikov, Kiselev,
& Jespersen 2018; Veraart, Novikov, & Fieremans, 2018; Novikov, Ver-
aart et al., 2018). Besides, WMTI parameterisation works in the case of
quite coherent axonal bundle with an orientation dispersion below 30°
(Fieremans et al., 2011). WMTI allows one to derive axonal water frac-
tion (AWF), extra-axonal axial and radial diffusivities (axEAD and
radEAD, respectively).

2.2.3 | Tract based spatial statistics

In order to evaluate and compare different QC approaches for the
derived diffusion maps, we applied TBSS (Smith et al., 2007). Initially,
all FA volumes were aligned to the FMR158_FA template, supplied by
FSL, using non-linear transformation implemented by FNIRT (Andersson
& Jenkinson, 2019). Next, a mean FA image across 18,600 subjects was
obtained and thinned in order to create mean FA skeleton. Afterwards,
each subject’s FA data are projected onto the mean FA skeleton, by fill-
ing the skeleton with FA values from the nearest relevant tract centre.
TBSS minimises confounding effects due to partial voluming and residual
misalignments originated from non-linear spatial transformations. For
each diffusion metric, we computed the individual skeleton projecting
the non-FA values onto the FA skeleton.

2.2.4 | QC model description

Our approach of image quality estimation originates from multi-
dimensional experiments in nuclear magnetic resonance spectroscopy
(Ernst, Bodenhausen, & Wokaun, 1987), when an additional dimension
allows one to resolve hidden resonance peaks. Natural parameter in
diffusion scalar metrics is an absolute value which either has its physi-
cal limitations, for example, FA and AWF lie between 0 and 1 and dif-
fusion kurtosis is limited by [0, 3] range (Tabesh, Jensen, Ardekani, &
Helpem, 2010; Veraart, Van Hecke, & Sijbers, 2011), or some region-
specific values from other sources, for example, free water diffusivity
in the brain equals to 3 μm²/ms. Thus, by applying a reasonable
threshold rule one can discard volumes with unfeasible values from
further analysis. However, since the values are typically averaged over
the volume (region of interest, skeleton, etc.), we still can expect “hid-
den” outliers with minimal influence on the averaged metric.

The general workflow of the proposed QC algorithm is
summarised in Figure 2. The workflow consists of five steps: first, we
estimate diffusion metrics for each subject in the diffusion space; sec-
ond, a normalisation step is performed in order to align each diffusion
map to Montreal Neurological Institute (MNI) space using FA map and
derived non-linear transformation; third, two QC parameters are esti-
mated for each subject: namely, averaged diffusion metric and SSIM.
SSIM values are estimated using the cohort mean diffusion metric as a
reference, mean diffusion values are obtained by averaging the scalar
maps over the TBSS skeleton; fourth, k-means approach for one clus-
ter allows us to obtain a distribution of the Euclidean distances for
each subject point to the cluster centroid; finally, the median distance
and empirically determined number of neighbours are used for the
density based clustering in order to identify possible outliers among
the derived diffusion metrics. As a result, the outlier exclusion is done
in the level of the whole brain volume.

We assume that SSIM allows us to spread image parameterisation
into the second dimension using three principal features: luminance,
contrast and structure as following (Wang et al., 2004),

\[
ssim(x, y) = \left( 1 + \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_2} \right) \cdot \left( 1 + \frac{2\sigma_x\sigma_y + c_3}{\sigma_x^2 + \sigma_y^2 + c_2} \right) \cdot \left( 1 + \frac{\sigma_{xy} + c_4}{\sigma_x\sigma_y + c_3} \right),
\]

where index x belongs to the evaluated map and y to the population
mean (reference) map for the given diffusion metric. \( \mu_{xy} \) are
the means of x and y, \( \sigma_{x,y} \) are the variances of x and y and \( \sigma_{xy} \) is the
covariance of x and y; constants \( c_{1,2,3} \) are the variables stabilising the
SSIM estimation, and $\alpha$, $\beta$, and $\gamma$ are the weights of three SSIM features. The stabilisation constants $c_1$, $c_2$ and $c_3$ are defined in Matlab as $c_4 = (0.01*L)^2$; $c_2 = (0.03*L)^2$; $c_3 = c_2/2$; with $L$ specified by a dynamic range value: $L = 1$ for images with $[0,1]$ scale, and 255 for others. By design, the SSIM metric is devoted to extract structural differences between the images, in contrast to the conventional approaches based on pixelwise error visualisation. Thus, SSIM is capable to identify differences between the information extracted from the scalar reference image and target image, similar to human visual perception. For the estimations, we used the $\text{ssim}$ function, implemented in Matlab. The SSIM values are estimated for each diffusion metric separately, such as FA, MD, MK and so on, due to different map structure of the diffusion metrics. In theory, SSIM estimator can be improved for some specific purposes (Charrier, Knoblauch, Maloney, Bovik, & Moorthy, 2012) or generalised (Brunet, Vrscay, & Wang, 2012). Nevertheless, original SSIM metric already proves its capability in medical image quality verification (Chow & Paramesran, 2016; Renieblas, Nogués, González, Gómez-Leon, & del Castillo, 2017; Vinding et al., 2017). The weights $\alpha$, $\beta$, and $\gamma$ allow one to emphasise the principle SSIM features in order to enhance a contrast between original image and target image, similar to human visual perception. For the estimations, we used the $\text{ssim}$ function, implemented in Matlab. The SSIM values are estimated for each diffusion metric separately, such as FA, MD, MK and so on, due to different map structure of the diffusion metrics. In theory, SSIM estimator can be improved for some specific purposes (Charrier, Knoblauch, Maloney, Bovik, & Moorthy, 2012) or generalised (Brunet, Vrscay, & Wang, 2012). Nevertheless, original SSIM metric already proves its capability in medical image quality verification (Chow & Paramesran, 2016; Renieblas, Nogués, González, Gómez-Leon, & del Castillo, 2017; Vinding et al., 2017). The weights $\alpha$, $\beta$, and $\gamma$ allow one to emphasise the principle SSIM features in order to enhance a contrast between original image and reference. While SSIM weights adjustment is still debated (Li, & Bovik, 2009), we empirically define the following weights $\alpha = .1$, $\beta = .1$ and $\gamma = 2$, in order to stretch a range of SSIM value components.

After the diffusion metric evaluation and normalisation of the scalar maps to the common space, we estimated the SSIM metrics for each diffusion map using averaged normalised diffusion metric as a reference image in SSIM. We performed outlier detection using the following two-step approach: first, we used $k$-means clustering (implemented as $\text{kmeans}$ Matlab function) to define one cluster based on squared Euclidean distances for (diffusion metric, SSIM) pairs (David & Vassilvitskii, 2007). Next, in order to introduce object density parameterisation, we used the median distance of the distance distribution (MDD) around the cluster centroid as a unit for neighbourhood radius in density-based spatial clusterisation algorithm with noise ($\text{dbscan}$, implemented as MATLAB function; Daszykowski, Walczak, & Massart, 2002; Ester, Kriegel, Sander, & Xu, 1996). The optional parameters in the $\text{dbscan}$ algorithm are number of objects in a neighbourhood of a central object and number of MDD units. We empirically set it to be equal to 10 and 7, respectively, in the tests. As such, the chosen parameters allow us to apply cluster density estimation independently from the original data distribution (see, e.g., Figure 3 and Kendall correlation coefficients between diffusion metrics and SSIM values, estimated using Matlab function $\text{corr}$).

As a frequently applied QC approach for comparison purpose, we applied a simple threshold approach of $3 SD$ from the mean diffusion value after regressing out main effects of age, sex and site.

### 2.3 Statistical analysis

In order to assess the effects of our proposed QC pipeline on the sensitivity of the diffusion metrics we tested for associations with age and sex using linear models as implemented in the Matlab function $\text{lmfit}$. In the subsample of 799 subjects (724 UKB subjects + 75 artificially distorted maps) we employed the following general linear model (GLM):

$$y = b_0 + b_1 \text{Age} + b_2 \text{Sex} + b_3 \text{Site},$$

where Age is given in years, Sex and site as a dichotomous variable. We computed specificity and sensitivity of the automated artefact correction. Sensitivity is defined as a ratio of True Positive/(True Positive + False Negative) and specificity as a ratio of True Negative/(True Negative + False Positive).

In the full UKB sample we employed the following models:

$$y = b_0 + b_1 \text{Age} + b_2 \text{Sex} + b_3 \text{Site} + (b_4 \text{Age}^2).$$

We computed root mean squared error (RMSE) and $R^2$ as proxies for goodness-of-fit. We compared coefficients between models (before and after discarding datasets flagged by our QC pipeline) using the R package $\text{cocor}$.
In order to assess normality of the residuals from the linear models we used QQ-plots (Aldor-Noiman, Brown, Buja, Rolke, & Stine, 2013; implemented as `qqplot` Matlab function) and Kolmogorov–Smirnov (KS) test with W critical value (Kolmogoroff, 1933; Smirnov, 1948). The W critical values have been used as indirect measures of normality of the residuals. The KS tests were implemented as MATLAB function `kstest`.

### 2.4 Machine learning for brain age gap estimation

We estimated the influence of outliers on ML based brain age prediction. The brain age gap (BAG) is defined as the difference between chronological and predicted age and has been proposed to reflect a sensitive imaging-derived phenotype (Kaufmann et al., 2019). For age prediction we applied two frequently used approaches. First, we employed linear model and multiple regressors (LMMR) defined as $Y = X\beta - \delta$, where $Y$ is the chronological age, $\beta$ is the regressor vector, $X$ is the matrix of brain features used for prediction, and $\delta$ is the BAG. The solution can be obtained by pseudo-inversion $X^+$ matrix (Smith et al., 2019). In order to improve the ML-training, we used 25% of eigenstates produced by the singular value decomposition replacing the $X$ matrix as recommended by Smith et al., 2019. In order to assess the influence of outliers on age prediction, a fixed number of 476 outliers, identified by the proposed QC approach over all diffusion metrics, was combined with varying samples of good-quality data, creating total training-sets of 1,000, 2000, 3,000, 4,000, 5,000, 7,500, 10,000, 12,500 and 15,000 subjects. The 476 outliers were manually added to each sample, leading to outlier percentages of 47.6, 23.8, 15.9, 11.9, 9.52, 6.35, 4.76, 3.81 and 3.17%, respectively. In all training sets, we kept the sex and site distribution identical. All training sets were selected from the whole UKB dataset. Thousand subjects not included in the training sets were selected as a test sample that was used in all runs. We performed the BAG estimations separately for the training sets with and without outliers, respectively. As criteria we used the Pearson correlations between chronological and predicted ages, and root mean squared errors estimated for the test sample.
2.5 | Simulated artefacts in small subsample

In order to verify our approach for detection of possible “badly” estimated scalar metrics and outliers we used a random subset of UKB data consisting of 724 subjects. Next, we manually introduced three types of image distortions to the evaluated diffusion scalar maps in MNI space. The first type (Type 1) is based on complete loss of \( N_1 \) random slices in the image volume. In our case, we set upper bound of \( N_1 \) to be equal 5. The second type of artefacts (Type 2) is based on value scaling of up to \( N_2 = 7 \) random slices. Scaling of the random slices can be performed in two ways as a division or multiplication of the diffusion values. To dilute scaled values between neighbouring slices we applied 3D Gaussian smoothing with 3mm\(^3\) kernel. The final type of distortions (Type 3) is based on residual misalignments along an image normalisation process. As a simple implementation of the residual misalignments we used rotation around superior–inferior axis with a random angle up to 5°. An example of original diffusion maps and three types of artificial distortions is presented in Figure 3. Finally, we added 25 volumes (10.4% of original data) of each type of distortions to four diffusion metrics: FA, MD, MK and AWF. In order to test influence of artefacts on the derived mean metric maps, we evaluated SSIM parameters of initial datasets using mean maps with (799 volumes) and without (724 volumes) outliers.

3 | RESULTS

3.1 | Sensitivity to simulated artefacts

In the randomly selected subset of UKB data we evaluated an influence of artificial outliers on the averaged diffusion metrics and estimated SSIM metrics. Resulting SSIM correlations are presented in Figure 3. High linear correlations (over .999) demonstrated that an introduction of outliers into data subsets did not influence on mean reference images and, consequently, on the SSIM evaluations. Supporting Information provide examples of the diffusion maps detected in the whole UKB sample with different types of distortions after data processing and scalar metric evaluation.

Figure 4 shows an application of the developed QC method to the subsample consisting of 799 subjects with artificially introduced outliers of the three types. The sensitivity of the QC method for the diffusion metrics is presented in Table 1. Briefly, based on AWF we detected all three types of introduced outliers. In the case of FA, we missed 10 (13%) outliers; in the case of MD, we missed 7 (9%) outliers; in the case of MK, we missed 2 (3%) outliers. In contrast, the QC approach based on data truncation beyond three SDs from the mean allowed us to detect for FA, only six outliers (69 outliers are missed, 92%); for MD, it detected only four outliers (71 outliers are missed, 95%); for MK, it detected six outliers (69 outliers are
missed, 92%); and for AWF, it detected four outliers (71 outliers are missed, 95%).

Figure 5 shows the various diffusion metrics plotted as a function of age and the corresponding linear fits based on the GLM. Briefly, the raw data and thresholding method yielded similar GLM parameters (see Table 2 for the intercept and slope values). Cocor function revealed no significant slope differences for any of the diffusion metrics between the raw and QC’ed data. Table 2 summarises the goodness-of-fit measures for the selected diffusion metrics for three datasets (raw data, thresholding QC and the developed QC method) and GLM parameters ($b_0$, intercept, and $b_1$, age slope). QQ plots and the W parameters from KS tests based on FA, MD, MK and AWF for three datasets (raw data, thresholding QC and our QC method) suggested that our proposed QC method yields the most “normal” residuals (Figure 6).

### 3.2 Effects of QC pipeline on the sensitivity to age, sex and scanner site

Figure 7 shows an application of the developed QC method to the UKB data with 18,608 subjects. As detailed above, we discarded datasets defined as outliers based on mean skeleton diffusion metrics and SSIM. The mean diffusion maps used as a reference for SSIM estimations are depicted in Supporting Information. Distributions of relevant diffusion metrics and demographics of the data

| Metrics | Type 1 (sensitivity/specificity) | Type 2 (sensitivity/specificity) | Type 3 (sensitivity/specificity) | Number of discarded volumes. False positive |
|---------|---------------------------------|---------------------------------|---------------------------------|---------------------------------------------|
| FA      | 2 (0.92/0.80)                   | 8 (0.68/0.80)                   | 0 (1/0.80)                      | 142                                         |
| MD      | 1 (0.96/0.91)                   | 6 (0.76/0.91)                   | 0 (1/0.91)                      | 69                                          |
| MK      | 0 (1/0.97)                      | 2 (0.92/0.97)                   | 0 (1/0.97)                      | 21                                          |
| AWF     | 0 (1/0.85)                      | 0 (1/0.85)                      | 0 (1/0.85)                      | 112                                         |

Table 1: Sensitivity and specificity of the proposed QC method for the diffusion metrics based on artificial distortions of three types (see Figure 3)
TABLE 2  Results of GLM $y = b_0 + b_1 \text{Age} + b_2 \text{Sex} + b_3 \text{Site}$ for four diffusion metrics using test sample of 724 subjects

| Metrics/statistics | Raw data No 799 | Threshold with 3 SD | Our QC method |
|--------------------|----------------|--------------------|---------------|
|                    | RMSE $R^2$     | NO RMSE $R^2$      | NO RMSE $R^2$ |
| FA                 | 0.0211 .112    | 792 0.0198 .137    | 592 0.0135 .1  |
| MD                 | 0.0327 .125    | 786 0.03 1.141     | 657 0.0223 .144 |
| MK                 | 0.0416 .0999   | 791 0.0391 .0985   | 705 0.0345 .105 |
| AWF                | 0.0152 .103    | 793 0.0145 .108    | 611 0.0109 .0629 |

|                | Intercept | Slope     | Intercept | Slope     | Intercept | Slope     |
|----------------|-----------|-----------|-----------|-----------|-----------|-----------|
| FA             | 0.5164 −9.80 $10^{-4}$ | 0.5197 $10^{-3}$ | 0.4990 $10^{-4}$ | −5.98 $10^{-4}$ |
| MD             | 0.8037 1.54 $10^{-3}$ | 0.8037 1.53 $10^{-3}$ | 0.8194 1.18 $10^{-3}$ |
| MK             | 1.1188 −1.55 $10^{-3}$ | 1.1154 −1.43 $10^{-3}$ | 1.1189 −1.38 $10^{-3}$ |
| AWF            | 0.4174 −6.29 $10^{-4}$ | 0.4176 −6.19 $10^{-4}$ | 0.4077 −3.80 $10^{-4}$ |

Note: RMSE is the root mean squared error; $R^2$ is the R-squared parameter; NO is the number of observations; SDs; Intercept is $b_0$; Slope is $b_1$.

FIGURE 6  QQ plots of the GLM residuals for three cases (see Figure 4): (a) raw data; (b) after thresholding by three SDs; (c) the proposed QC method. Values $W$ are the critical numbers of Kolmogorov–Smirnov (KS) test for a normality. In all cases, KS test did not reveal that the residuals are normally distributed.
**FIGURE 7** An application of QC method to UKB data. Mean diffusion maps are presented as a reference for the SSIM estimations. The red circles are identified outliers; the blue circles are the filtered data. SSIM, FA, MK and AWF are unit-less values, MD is in $\mu m^2/ms$. The dashed black lines mark the boundaries in three SDs from the mean value.

**FIGURE 8** The results of GLM age-diffusion correlations with linear age term (red line) and quadratic age term (black line). The plots marked as “All” consists of all raw data ($n = 18,608$); the plots marked as “QC” consists of data passed through the QC filtration ($n = 18,132$). Intervals of confidence (CI 95%) are presented as dashed line in all cases. FA, MK and AWF are unit-less values, MD is in $\mu m^2/ms$. 
defined as outliers are presented in Supporting Information. A higher number of outliers were identified from the Cheadle \( (n = 396; 3\%) \) site compared to the Newcastle \( (n = 78; 1\%) \) site, and both sex (39% of women) and age distribution (58.55/7.74 years) of the outlier data did not diverge substantially from the distributions in the total sample. For illustration, we depicted the boundaries of \( 3 \times SD \) from the mean cohort values for each diffusion metric.

Table 3 presents age-related trajectories (linear and quadratic fits) for four diffusion metrics with the detected outliers included or excluded. The summary statistics are summarised in Table 3. All other metrics are shown in the Supporting Information. For illustration, we depicted the boundaries of \( 3 \times SD \) from the mean cohort values for each diffusion metric.

3.3 | Effects of QC pipeline on the ML BAG estimations

Figure 10 shows the trajectories of RMSE and correlations between chronological and predicted ages for the two ML algorithms. For statistics and cross validation of the BAG results, we repeated model training 100 times, randomly choosing the training samples from whole UKB data. Briefly, in the case of QC filtered data model performance increased only moderately with sample size, with RMSE of the XGBoost algorithm suggesting only minor effects. In contrast, the training sets with outliers demonstrated strong dependence of the chronological and predicted age correlations and RMSE on the percentage of outliers in the training sample, in particular, for the LMMR approach. For both ML algorithms, increasing training sample size decreased RMSE in the test set.

### DISCUSSION

Advanced dMRI offers sensitive measures of brain tissue micro- and macrostructural architecture and integrity, with large potential for the basic and clinical neurosciences. With the surge of large-scale clinical and population-based efforts acquiring dMRI data from thousands of individuals, there is an increasing need to develop computationally
efficient pipelines for quality assessment and identification of poor quality data among derived diffusion scalar metrics. The proposed QC method based on 2D data representation exploiting similarity metrics and data density features enables an efficient evaluation of data quality after estimation of diffusion scalar metrics. In a subsample \( n = 724 \) plus 75 artificial outliers, our semi-automated artefact detection based on similarity metrics yielded high sensitivity and specificity and the residuals from linear age fits in the full sample \( n = 18,608 \) resulted in more normal residuals after compared to before discarding flagged datasets. Additionally, the QC pipeline improved brain-age prediction using ML by mitigating the influence of outliers in the training set.

By default, the harmonised validated raw diffusion data allow one to derive accurate scalar metrics. However, a quality evaluation of the processed diffusion maps is still an open question in big data analysis. Many efforts have been made to develop accurate QC and harmonisation procedures on raw diffusion weighted data (Fortin et al., 2017; Mirzaalian et al., 2018). Nevertheless, derived diffusion metrics from DTI or DKI may still deviate from expected range, for example, due to remaining artefacts and numerical misestimations (see Supporting Information for examples of the distorted diffusion maps). Despite improved post-processing algorithms (Ades-Aron et al., 2018) for raw diffusion data, there is no consensus yet about a unified pipeline for diffusion data, for example, noise correction methods are regularly revised (Muckley et al., 2021), Gibbs ringing artefacts can remain in the images due to different origins such as a partial Fourier (Muckley et al., 2021), frequency drift effect (Vos et al., 2016) can bias the estimations, in particular in the case of advanced dMRI protocols, and diffusion gradient non-linearity correction (Rudrapatna, Parker, Roberts, & Jones, 2020) might be important as well. Notably, a number of artefacts in the scalar diffusion maps could be minimised by applying a state-of-the-art algorithms such as, for example, eddy_gpus, if a computational facility allows that. The simple considerations of 2D representations of the averaged diffusion

**FIGURE 9** QQ-plots of the GLM residuals for four diffusion metrics. “Original, Linear” means the all data and GLM with linear age term; “Original, Quadratic” means the all data and GLM with quadratic age term; “QC, Linear” means the QC filtered data and GLM with linear age term; “QC, Quadratic” means the QC filtered data and GLM with quadratic age term.
metric and SSIM values are an advantage of the developed QC method. This allows us to take into account frequently applied measures in large-scale studies, that is, diffusion metrics averaged across a region of interest or the entire TBSS skeleton and the structural similarity based on the intensity, contrast and structure of the scalar map in relation to a reference map. As an improvement of the developed approach, one could generalise SSIM parameter for the scalar maps belonging to the same diffusion approach. A combination of various diffusion contrasts could improve the QC efficacy and reduce a computation time. Our simulations revealed that our method is capable of identifying image artefacts with different origins with high sensitivity and specificity. Notably, our approach allows one to reveal different artefacts originated either from the corrupted diffusion maps or caused by the not accurate warping procedure to MNI space. In the case of map misalignments, the original diffusion maps in the diffusion space still can be used in the native space, for example, for a tractography but should be processed separately in the following group analysis. This is particularly valuable in the context of large-scale studies, where manual QC is not feasible and when a quantitative estimate of structural similarity is needed. Whereas our direct comparison of slopes did not reveal a significant effect of the QC procedure on the estimated age-associations, the linear models based on QC’ed data yielded evidence of improved model fits in terms of the distributions of the model residuals compared to models based on the non-QC’ed data.

We found evidence of improved ML based age prediction when limiting the number of noisy datasets in the training set. In general, larger training sets are expected to increase accuracy of brain age prediction (see, e.g., Kaufmann et al., 2019). However, in practice, the number of accessible data is usually limited. Thus, it is very important to know how different amounts of undetected outliers in the training set could affect the prediction accuracy in an independent test set. Our results demonstrated that a higher portion of outliers in the training set influenced the prediction accuracy in the test set. Surprisingly, for XGBoost, in contrast to the RMSE, the correlation between predicted and chronological age did not increase much with increasing training set size. For LMMR the correlation coefficients increased in accordance with increased sample size. In both instances, however,
the proportion of bad datasets in the training set influenced the prediction accuracy in the test set, with markedly improved prediction with lower proportion of noisy data.

In summary, although an overall beneficial effect of removing poor quality datasets results is not surprising, our results serve as relevant demonstrations of the importance of QC in the context of large-scale studies. It should also be noted that all datasets included in the current analysis have been checked and approved by the initial U.K. Biobank QC procedures (Alfaro-Almagro et al., 2018), and the reported effects of noise removal on age-associations are likely to represent lower-bound effects compared to a scenario with no initial QC procedures. In general, whereas minimising noise is a universal aim, the direct effects and value of QC will vary between studies and applications. As a relevant verification of the proposed QC approach, we plan to apply the same procedure to other imaging biobanks such as Adolescent Brain Cognitive Development study, conceived and funded by the National Institutes of Health, United States.

Conclusively, in the case of big data, automated, efficient and reliable approaches for evaluating the scalar diffusion metrics prior to statistical analysis are needed. Our results suggest that our proposed method is suitable as a complementary test of the estimated diffusion data to increase sensitivity of conventional diffusion scalar metrics.

ACKNOWLEDGEMENT

This work was funded by the Research Council of Norway (249795 and 276082). This research has been conducted using the TSD (Tjeneste for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT-Department (USIT). Computations were also performed on resources provided by UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway. IIM thanks Dr. Viljami Sairanen for his help with FSL and slurm queue scripting.

DATA AVAILABILITY STATEMENT

All used data are accessible through U.K. Biobank service.

ORCID

Ivan I. Maximov https://orcid.org/0000-0001-6319-6774
Lars T. Westlye https://orcid.org/0000-0001-8644-956X

REFERENCES

Ades-Aron, B., Veraart, J., Kochunov, P., McGuire, S., Sherman, P., Kellner, E., ... Fieremans, E. (2018). Evaluation of the accuracy and precision of the diffusion parameter ESTimation with Gibbs and NoisE removal pipeline. *NeuroImage*, *183*, 532–543. https://doi.org/10.1016/j.neuroimage.2018.07.066

Aldor-Noiman, S., Brown, L. D., Buja, A., Rolke, W., & Stine, R. A. (2013). The power to see: A new graphical test of normality. *The American Statistician*, *67*(4), 249–260. https://doi.org/10.1080/00031305.2013.847865

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffianti, L., Douaud, G., ... Smith, S. M. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*(February), 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Andersson, J., Jenkinson, M., & Smith, S. (2007). Non-linear registration aka spatial normalisation FMRIB technical report TR07JA2. Retrieved from https://www.fmrib.ox.ac.uk/datasets/techrep/tr07ja2/tr07ja2.pdf.

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

Basser, P. J., Mattiello, J., & Lebihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin Echo. *Journal of Magnetic Resonance, Series B*, *103*(3), 247–254. https://doi.org/10.1016/0167-8105(94)90103-7

Bastiani, M., Cottaar, M., Fitzgibbon, S. P., Surí, S., Alfaro-Almagro, F., Sotiropoulos, S. N., ... Andersson, J. L. R. (2019). Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction. *NeuroImage*, *184*(January), 801–812. https://doi.org/10.1016/j.neuroimage.2018.09.073

Brunet, D., Vrcay, E. R., & Wang, Z. (2012). On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, *21*(4), 1488–1499. https://doi.org/10.1109/TIP.2011.2173206

Charrier, C., Knoblauch, K., Maloney, L. T., Bovik, A. C., & Moorthy, A. K. (2012). Optimizing Multiscale SSIM for Compression via MLDS. *IEEE Transactions on Image Processing*, *21*(12), 4682–4694. https://doi.org/10.1109/TIP.2012.2210723

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining—KDD ’16* (pp. 785–794). San Francisco, CA: ACM Press. https://doi.org/10.1145/2939672.2939785

Chow, L. S., & Paramesran, R. (2016). Review of medical image quality assessment. *Biomedical Signal Processing and Control*, *39*, 1–11. https://doi.org/10.1016/j.bspc.2016.02.006

Daszykowski, M., Walczak, B., & Massart, D. L. (2002). Looking for natural patterns in analytical data. 2. Tracing local density with OPTICS. *Journal of Chemical Information and Computer Sciences*, *42*(3), 500–507. https://doi.org/10.1021/ci010384s

David, A., & Vassilvitskii, S. (2007). K-Means++: The advantages of careful seeding. 1027–1035.

David, S., Mesri, H. Y., Viergever, M. A., & Leemans, A. (2019). Statistical significance in DTI group analyses: How the choice of the estimator can inflate effect sizes. *BioRxiv* 755140. https://doi.org/10.1101/755140

de Lange, A.-M. G., Barth, C., Kaufmann, T., Maxi moment, I. L., van der Meer, D., Agartz, I., & Westlye, L. T. (2020). Women’s brain aging: Effects of sex-hormone exposure, pregnancies, and genetic risk for Alzheimer’s disease. *Human Brain Mapping*, *41*(18), 5141–5150. https://doi.org/10.1002/hbm.25180

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Ernst, R. R., Bodenhausen, G., & Wokaun, A. (1987). Principles of nuclear magnetic resonance in one and two dimensions. In *The international series of monographs on chemistry* 14. Oxford [Oxfordshire]: New York: Clarendon Press; Oxford University Press.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96)*. AAAI Press, 226–231.
