HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2017 January 15.

Published in final edited form as:

Neuroimage. 2016 January 15; 125: 903–919. doi:10.1016/j.neuroimage.2015.10.068.

The Impact of Quality Assurance Assessment on Diffusion Tensor Imaging Outcomes in a Large-Scale Population-Based Cohort

David R. Roalf[a], **Megan Quarmley**[a], **Mark A. Elliott**[b], **Theodore D. Satterthwaite**[a], **Simon N. Vandekar**[a,e], **Kosha Ruparel**[a], **Efstathios D. Gennatas**[a], **Monica E. Calkins**[a], **Tyler M. Moore**[a], **Ryan Hopson**[a], **Karthik Prabhakaran**[a], **Chad T. Jackson**[a], **Ragini Verma**[b,c], **Hakon Hakonarson**[d], **Ruben C. Gur**[a,b], and **Raquel E. Gur**[a,b]

[a]Neuropsychiatry Section, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[b]Department of Radiology, University of Pennsylvania Perelman School of Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[c]Section of Biomedical Image Analysis, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[d]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[e]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia PA 19104, USA

Abstract

Background—Diffusion tensor imaging (DTI) is applied in investigation of brain biomarkers for neurodevelopmental and neurodegenerative disorders. However, the quality of DTI measurements, like other neuroimaging techniques, is susceptible to several confounding factors (e.g. motion, eddy currents), which have only recently come under scrutiny. These confounds are especially relevant in adolescent samples where data quality may be compromised in ways that confound interpretation of maturation parameters. The current study aims to leverage DTI data from the Philadelphia Neurodevelopmental Cohort (PNC), a sample of 1,601 youths ages of 8–21 who underwent neuroimaging, to: 1) establish quality assurance (QA) metrics for the automatic identification of poor DTI image quality; 2) examine the performance of these QA measures in an external validation sample; 3) document the influence of data quality on developmental patterns of typical DTI metrics.

*Please address correspondence to: David R. Roalf, Ph.D., Department of Psychiatry, Neuropsychiatry Section, Brain Behavior Laboratory, 10th Floor, Gates Building, Hospital of the University of Pennsylvania, Philadelphia, PA 19104, roalf@upenn.edu.

CONFLICT OF INTERESTS: The authors declare no competing financial interests.

Methods—All diffusion-weighted images were acquired on the same scanner. Visual QA was performed on all subjects completing DTI; images were manually categorized as Poor, Good, or Excellent. Four image quality metrics were automatically computed and used to predict manual QA status: Mean voxel intensity outlier count (MEANVOX), Maximum voxel intensity outlier count (MAXVOX), mean relative motion (MOTION) and temporal signal-to-noise ratio (TSNR). Classification accuracy for each metric was calculated as the area under the receiver-operating characteristic curve (AUC). A threshold was generated for each measure that best differentiated visual QA status and applied in a validation sample. The effects of data quality on sensitivity to expected age effects in this developmental sample were then investigated using the traditional MRI diffusion metrics: fractional anisotropy (FA) and mean diffusivity (MD). Finally, our method of QA is compared to DTIPrep.

Results—TSNR (AUC=0.94) best differentiated Poor data from Good and Excellent data. MAXVOX (AUC=0.88) best differentiated Good from Excellent DTI data. At the optimal threshold, 88% of Poor data and 91% Good/Excellent data were correctly identified. Use of these thresholds on a validation dataset (n=374) indicated high accuracy. In the validation sample 83% of Poor data and 94% of Excellent data was identified using thresholds derived from the training sample. Both FA and MD were affected by the inclusion of poor data in an analysis of age, sex and race in a matched comparison sample. In addition, we show that the inclusion of poor data results in significant attenuation of the correlation between diffusion metrics (FA and MD) and age during a critical neurodevelopmental period. We find higher correspondence between our QA method and DTIPrep for Poor data, but we find our method to be more robust for apparently high-quality images.

Conclusion—Automated QA of DTI can facilitate large-scale, high-throughput quality assurance by reliably identifying both scanner and subject induced imaging artifacts. The results present a practical example of the confounding effects of artifacts on DTI analysis in a large population-based sample, and suggest that estimates of data quality should not only be reported but also accounted for in data analysis, especially in studies of development.

Keywords

Diffusion tensor imaging; automated quality assurance; motion; adolescence; brain maturation

1.0 Introduction

Diffusion tensor imaging (DTI) is an important magnetic resonance imaging (MRI) technique in the investigation and identification of brain biomarkers in typical neurodevelopment, cognitive aging and neuropsychiatric syndromes. DTI is based on the quantification of the random Brownian motion of protons, which enables the measurement of the spatial organization of brain tissue (Basser et al., 1994; Le Bihan et al., 2001). Specifically, DTI provides contrasts that are sensitive to intra-voxel white matter microstructure (Basser and Pajevic, 2000) and produces results that are consistent with the major white matter pathways detailed in animal models and in retinotopic studies in the human brain (Conturo et al., 1999; Le Bihan, 2003). An extensive DTI literature details findings in normal development (e.g. (Ladouceur et al., 2012; Lenroot and Giedd, 2010; Oishi et al., 2013; Yoshida et al., 2013) and neuropsychiatric conditions, including

schizophrenia (Roalf et al., 2015; Wheeler and Voineskos, 2014), autism (Konrad and Eickhoff, 2010; Travers et al., 2012) and Alzheimer's disease (Radanovic et al., 2013; Zhang et al., 2014). Moreover, large consortia, such as the Human Connectome Project (Van Essen et al., 2012) and the Alzheimer's Disease Neuroimaging Initiative (Jack et al., 2010) rely on DTI data as a major outcome measure. However, DTI is not without practical challenges that affect the reliability and reproducibility of results (Le Bihan et al., 2006).

Neuroimaging data confounds, in particular head motion, are quite pertinent in human samples (Liu et al., 2015; Power et al., 2012), especially children (Yoshida et al., 2013) and adolescents (Satterthwaite et al., 2012). For example, the confounding influence of head motion on resting-state functional connectivity has received substantial attention (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). Similar effects are evident in structural MRI (Reuter et al., 2015) and pediatric DTI samples (Lauzon et al., 2013; Yoshida et al., 2013). DTI measurements, in general, are reliable as they are insensitive to B_1 errors, however, DTI is strongly dependent on gradient calibration and errors in gradient amplitude, direction and linearity, can contribute to inaccurate measurement (Conturo et al., 1995). Nonetheless, the influence of data quality using DTI is understudied and often ignored. Beyond head motion, the quality of DTI measurements is susceptible to several confounds including eddy currents, scanner artifacts (e.g. noise spikes) and susceptibility artifacts (Anderson, 2001; Bastin et al., 1998; Skare et al., 2000), which have come under scrutiny (Heim et al., 2004; Jones et al., 2013; Jones et al.; Lauzon et al., 2013; Owens et al., 2012; Tournier et al., 2011; Yendiki et al., 2014), and are the focus of several new methods seeking to mitigate their impact (Li et al., 2014; Li et al., 2013; Oguz et al., 2014). These confounds likely contribute to inaccuracies in the tensor fitting of DTI data (Le Bihan et al., 2006). For example, head-motion was found to induce group differences between autistic and typically developing children in DTI (Yendiki et al., 2014). Most importantly, the use of head-motion as a nuisance regressor during statistical modeling reduced this effect. Yet, most developmental and clinical studies using DTI fail to report procedures for quality control and its impact on results. It is likely that unaccounted for artifacts result in suboptimal tensor estimation and thus may negatively influence commonly derived DTI metrics, such as fractional anisotropy, mean diffusivity, and estimates of tractography. In addition, because data quality is often systematically related to a phenotype of interest (e.g. age, diagnosis, cognition, symptom severity) and that data quality is inherently subject dependent (e.g. correlation between age and motion), low quality data has the potential to obscure the presence of real effects or produce spurious associations with study phenotypes.

Despite such dangers, automated measures for quality assurance (QA) of DTI data remain limited. Manual inspection of multivolume DTI data is time consuming, subjective and potentially susceptible to operator bias, and translates poorly to large-scale imaging studies. Studies of noise in DTI provide a useful framework for identifying how such noise affects diffusion properties (Ding et al., 2005; Farrell et al., 2007; Hasan, 2007; Skare et al., 2000). Several recent studies indicate promise for implementing automatically derived quality assurance metrics that reduce the amount of manual QA effort, including measures of signal-to-noise and the use of outlier detection, to quantify data quality prior to image processing (Lauzon et al., 2013; Li et al., 2014; Li et al., 2013; Oguz et al., 2014). However, much of this work has used relatively small samples or simulated data, and none have

focused primarily on a neurodevelopment sample (although Lauzon et al., 2013 present data in a large pediatric sample). Finally, there is lack of corroboration of derived metrics in a validation sample.

The overall goal of the current study is to determine which image QA metrics are most reliable in the automatic detection of poor DTI data. We manually evaluate over 1,500 DTI data sets from the Philadelphia Neurodevelopmental Cohort (PNC; (Gur et al., 2014; Satterthwaite et al., 2014), and automatically derive QA measures. This approach will be useful in the current cross-sectional sample, in concurrent or longitudinal studies, and generalizable to most DTI studies. Importantly, all DTI data in the PNC was acquired within a 30-month period using the same MRI scanner, head-coil and DTI protocol. In addition, 25% of the sample returned for a follow-up DTI scan approximately two years later, thus providing a unique validation sample. Our goals are: 1) leverage DTI data from the PNC, a sample of 1,601 youth between the age of 8–21 who underwent neuroimaging, to determine automated quality assurance metrics that will aid in the automatic identification of poor DTI image quality; 2) test these QA measures in a follow-up sample; and, 3) determine the influence of data quality on typical DTI metrics (e.g. FA and MD), 4) measure changes introduced by including poor data in the correlations between FA/MD and age and 5) compare our QA processes to a previous published DTI QA tool, DTIPrep (Oguz et al., 2014). As described below, results indicate automated QA of DTI can facilitate large-scale, high-throughput analysis by reliably identifying poor quality data and systematically improving data fidelity.

2.0 Materials & Methods

2.1 Participants

2.1.1 Initial sample—All participants included in this study were enrolled in the PNC (Calkins et al., 2015; Calkins et al., 2014; Gur et al., 2014; Satterthwaite et al., 2014). The PNC is a large community-based epidemiological sample of 9,498 youths, aged 8–21, who underwent clinical and cognitive evaluations. A subset of 1,000 subjects received multimodal neuroimaging as part of the initial PNC project. An additional 601 individuals underwent the identical neuroimaging protocol as part of extension of the PNC. Data from 244 individuals was considered unusable (Figure 1). Accordingly, 1,357 individuals comprise the initial sample that received the same neuroimaging protocol (Table 1). These data were acquired between 2009 and 2012. A description of the PNC is available in: http://www.med.upenn.edu/bbl/projects/pnc/PhiladelphiaNeurodevelopmentalCohort.shtml and the data is available from the National Institutes of Health - dbGaP (http://www.ncbi.nlm.nih.gov/gap).

2.1.2 Validation sample—Four hundred and four individuals (Table 2) returned approximately two years later and underwent the same neuroimaging procedures. Of note, these individuals were selected to return based upon successful completion and high data fidelity of a structural scan during the initial study. DTI quality during the initial study was not a factor in enrollment for follow-up. Thirty individuals did not complete a follow-up DTI scan. Thus, the final sample was 374. These data were collected between 2012 and 2013.

All enrolled subjects provided informed consent at each visit, or for minors informed assent in addition to parental or guardian consent. The Institutional Review Boards of the University of Pennsylvania and Children's Hospital of Philadelphia approved all procedures.

2.1.3 Clinical assessment of the sample—Detailed clinical information on the PNC is published (Calkins et al., 2015; Calkins et al., 2014; Gur et al., 2014). While clinical diagnosis is not of interest in the current study, it is imperative that clinical diagnosis be considered when discerning the contributors to neuroimaging artifact. It is possible that individuals with significant clinical symptomatology may have more difficulty completing hour-long neuroimaging sessions than healthy individuals. To determine if clinical features affect quality of assessment procedures, baseline psychopathology is analyzed as a general psychopathology factor score reflecting levels of symptoms across psychopathology domains (Calkins et al., in prep).

2.2 Diffusion Tensor Imaging

2.2.1 Diffusion Weighted Imaging Acquisition—Recruitment information and detailed neuroimaging scan parameters for all sequences have been described (Satterthwaite et al., 2014). Briefly, all MRI scans were acquired on the same 3T Siemens Tim Trio whole-body scanner and 32-channel head coil at the Hospital of the University of Pennsylvania. DTI scans were obtained using a twice-refocused spin-echo (TRSE) single-shot EPI sequence (TR=8100 msec, TE=82 mesc, FOV=240mm^2/240mm^2; Matrix= RL: 128/AP: 128/Slices:70, in-plane resolution (x & y) 1.875mm^2; slice thickness=2 mm, gap=0; FlipAngle=90°/180° /180°, volumes=71, GRAPPA factor =3, bandwidth=2170 Hz/pixel, PE direction=AP). The sequence employs a four-lobed diffusion encoding gradient scheme combined with a 90–180–180 spin-echo sequence designed to minimize eddy-current artifacts (Reese et al., 2003). The complete sequence consisted of 64 diffusion-weighted directions with b = 1000 s/mm^2 and 7 interspersed scans where b = 0 s/mm^2 (Satterthwaite et al., 2014). Scan time was approximately 11 minutes. The imaging volume was prescribed in axial orientation covering the entire cerebrum with the topmost slice just superior to the apex of the brain. The diffusion encoding orientations are provided in Table 4.

In addition to the DTI scan a map of the main magnetic field (i.e. B0) was derived from a double-echo, gradient-recalled echo (GRE) sequence (Satterthwaite et al., 2014). Both magnitude and phase images were selected for image reconstruction since it is the phase signal that contains information about the magnetic field. A user-modified version of the multi-echo GRE sequence that enabled advanced shimming feature was used instead of the Siemens product B0 mapping sequence as this product was not available at study initiation. This procedure performs multiple passes of the automated shim current optimization, resulting in improved magnetic field homogeneity across the cerebrum. The field-of-view of this scan was chosen to be larger than that of the DTI scans in order to obtain a field map that covered all of the volumes of interest across all functional MRI and DTI sequences obtained during the imaging session.

2.2.2 Diffusion Tensor Imaging Quality Assurance (QA)—DTI data QA occurred in three steps. Initial visual QA was done to ensure data fidelity. Thus, individuals with

structural abnormalities were excluded. In addition, any data collected without the default scan parameters was excluded. Data passing initial QA were passed to manual QA, where a trained analyst inspected each DTI series in detail. In parallel, an automated QA process was run on each DTI series to extract QA metrics. These steps are detailed below.

2.2.3 Initial QA—Initial quality assurance (n=1601), which targeted gross abnormalities or deviations in protocol, is documented in Figure 1. As the DTI sequence was obtained toward the end of the neuroimaging session, this scan was sometimes not completed (n=162, 10%) or terminated prematurely due to time constraints (n=20, 1%). Individual MRI sessions with incidental findings (n=20, 1%) or a failure to collect a corresponding field map (n=22, 1%) were excluded from further analysis. Finally, alterations to the DTI acquisition parameters, such as a change in the echo time, also resulted in exclusion (n=20, 1%). After these exclusions 1,357 DTI scans were available for manual inspection.

2.2.4 Visual QA—A trained analyst visually inspected all 71 volumes for each DTI scan. Any volume that appeared to contain artifact was recorded. Two main types of artifacts were observed. Volumes failed QA due to signal dropout likely caused by the interaction of subject motion and diffusion encoding. Less common was image striping, which was likely caused by defective gradient performance (Figure 2). This is likely related to mechanical vibrations at the interface of the gradient cables and the magnet bore (Satterthwaite et al., 2014). Individual datasets were scored based on the number of volumes that included artifact. This scoring was based on previous work detailing influence of removing image volumes when estimating the diffusion tensor (Chen et al., 2015; Jones and Basser, 2004). Data was considered: 1) “Poor” if more than 14 (20%) volumes contained artifact; 2) “Good”, if one or more, but less than 15 volumes contained artifact; and 3) “Excellent”, if no volumes were flagged as containing artifact. This procedure was followed for both the initial and follow-up sample; a process that took approximately three months to complete by a trained reviewer.

2.2.5 Automated QA—Four image quality metrics were automatically computed from each DTI scan: 1) mean voxel outlier count (MEANVOX), 2) maximum voxel outlier count (MAXVOX), 3) mean relative motion (MOTION) and 4) temporal signal-to-noise ratio (TSNR). This automated QA script is freely available (cmroi.med.upenn.edu/QAscripts) and runs on NIFTI file format data.

2.2.5.1 Voxel intensity: The AFNI (Cox, 1996) tool 3dToutcount was used to determine the number of intensity outliers, at each time point in the DTI series. This tool estimates mean voxel intensity and the median absolute deviation (MAD) for each volume (i.e. time point) after brain masking. An outlier is determined to be any time point exceeding a threshold, as defined in 3dToutcount program, in units of MAD from the voxel mean minus its trend (Cox, 1996). The threshold is defined within 3dToutcount by iteratively determining if a point is ‘far away’ from the trend. ‘Far’ is determined using the following formula:

$$\text{alpha} * \text{sqrt}(\text{PI}/2) * \text{MAD}, \text{where}$$
$$\text{alpha}=\text{qginv}(0.001/\text{N}), \text{where}$$
$$\text{N}=64\ (\text{number of diffusion weighted volumes})$$

Thus, the output of this method is a single time series of outlier counts per volume. From these values the mean and maximum outlier count values are calculated. Only the 64 b=1000 s/mm^2 were utilized in this calculation. This approach not only provides valuable QA information but also specifies particular volumes that should be scrutinized.

2.2.5.2 Motion: The primary measurement of in-scanner head motion was mean relative volume-to-volume displacement as determined by rigid-body motion correction. This metric is commonly used in fMRI studies to measure head motion (Satterthwaite et al., 2012; Van Dijk et al., 2012). This standard measure summarizes total volume-to-volume translation and rotation across all three axes (Jenkinson et al., 2002). However, given the difficulty of estimating motion using low-intensity diffusion weighted images, motion was estimated from the high quality b=0 images (n=7) that were interspersed throughout DTI acquisition.

2.2.5.3 Average temporal signal-to-noise ratio: The temporal signal-to-noise ratio was estimated for each brain voxel using only the 64 b=1000 s/mm^2 DTI volumes. The voxel-wise SNR was calculated from the mean and standard deviation of each voxel's intensity, after brain masking and motion correction, was measured. Subsequently, the average of all brain voxel temporal SNR was calculated to report a single metric of overall scan SNR.

2.3 General DTI image processing

2.3.1 Preprocessing—Image processing steps and the tools used for each step are displayed in Figure 3. Briefly, DTI data were merged into a single series and QA procedures were run. The skull was removed by generating a brain mask for each subject by registering a binary mask of a standard image (FMRIB58_FA) to each subject's brain using FLIRT (Smith, 2002). When necessary, manual adjustments were made to this mask. Next, eddy currents and movement were estimated and corrected using FSL's eddy tool (Woolrich et al., 2001). Eddy is an improvement upon the typical eddy/motion correction used as part of FSL's Diffusion Tool Box (Behrens et al., 2003). This tool simultaneously models the effects of diffusion eddy current and head movement on DTI images in order to reduce the amount of resampling. The intended use for eddy is with high-direction, high b-value data that encompass the whole sphere, but it can be used for more common DTI data. Next, the diffusion gradient vectors were rotated to adjust for motion using the 6-parameter motion output generated from eddy (Smith, 2002). Then, the field map was estimated and distortion correction was applied to the DTI data using FSL's FUGUE (Satterthwaite et al., 2014). Finally, the diffusion tensor was modeled and metrics (FA and MD) were estimated at each voxel using FSL's DTIFIT.

2.3.2 Post-processing—Registration from native space to a template space was completed using DTI-TK (Zhang et al., 2014; Zhang et al., 2006). First, the FA and MD output of DTIFIT was converted to DTI-TK format. Next, a study specific template was

generated from the tensor volumes using 14 representative diffusion data sets that were considered "Excellent" after manual QA. One individual from each of the 14 ages (age range 8–21) was randomly selected. These 14 DTI volumes were averaged together to create an initial template. Next, data from the 14 subjects were registered to this template in an iterative manner. Unlike standard intensity-based registration algorithms, this process utilizes the full tensor information in an attempt to best align the underlying white matter tracts using iterations of rigid, affine and diffeomorphic registration leading to the generation of a successively refined template. Ultimately, one high-resolution refined template was created and used for registration of the remaining diffusion datasets. All FA and MD maps were then registered (rigid, affine, diffeomorphic) to the high-resolution study-specific template using DTI-TK. These maps were then used for group comparisons. Two subjects failed the registration process and were excluded from further analysis. Standard regions of interest (ROI; ICBM-JHU White Matter Tracts; Harvard-Oxford Atlas) were registered from MNI152 space to the study-specific template using ANTs registration (Avants et al., 2011). Mean FA and MD were extracted from these ROIs.

2.4 Demographically matched sub-sample

To evaluate the influence of the inclusion of suboptimal DTI data on FA and MD we created two age, sex and race matched samples (n=146, each) from the good and excellent data to match the individuals in the poor data group using an optimal matching algorithm written in SAS (SAS, Institute, Carey, NC, USA). The use of matched subsamples was necessary given that individuals with Poor data tended to be male and younger, while individuals with excellent data tended to be older and female. Race and psychopathology score were also included as an additional matching variable. The characteristics of the matched sample are presented in Table 5.

2.5 Comparison of QA methodology with DTIPrep

In order to further validate our QA methodology, we compared our QA approach to a newly available open-source program-- DTIPrep (Oguz et al., 2014). DTIPrep is a dedicated diffusion MRI (dMRI) quality control program designed to "identify and correct all common, known dMRI artifacts" (Oguz et al., 2014); pg 4). DTIPrep is a comprehensive tool that offers the following QA methodology: 1) dMRI protocol validation, 2) diffusion information validation, 3) interslice brightness artifact detection, 4) interlaced artifact detection ("venetian blind effect") on a slice-wise basis, 5) iterative averaging of non-weighted images, 6) motion and eddy current correction, including residual motion detection and 7) reconstruction of usable DTI data.

DTIPrep was performed on the matched sample (n=438) using the default parameters based upon our DTI acquisition. The tolerance threshold for 'failed' volumes/gradients was set at 20% of the total number of volumes (14/71). This tolerance level was chosen in accordance with the definition of "Bad" data in our visual QA. Only the steps noted above were included in the automated DTIPrep QA analysis.

2.6 Statistical Analysis

Demographic characteristics were compared across manual QA groups using Pearson χ^2 or one-way ANOVAs with post-hoc t-tests (Bonferroni corrected). To determine the ability of automated QA measures to correctly classify problematic data identified in manual QA we performed a two-stage analysis. In Stage 1, we were concerned with identifying the diffusion data that should be excluded from further analysis (poor data), thus two groups were created from the initial sample: Group 1 was comprised of the "Poor" data (n=146)[1] and the remainder of individuals ("Good" + "Excellent") comprised Group 2 (n=1,208). In Stage 2, we were concerned with identifying the diffusion data that was excellent from the data that might need some manual intervention to be usable. Here, Group 1 was comprised of "Good" data, while Group 2 was comprised of the data identified as "Excellent". The accuracy for each automated measure (MAXVOX, MEANVOX, MOTION, TSNR) was calculated as the area under the receiver operating characteristic (ROC) curve (AUC). The logistic ROC analysis used a 10-fold cross-validation approach to estimate AUC and optimal cut-off score. Larger AUC values indicate more accurate classification of participants. A cut-off score for each measure that best differentiated manual QA group was determined using the Youden Index (Youden, 1950), which maximizes the tradeoffs between sensitivity and specificity. The classification accuracy (probability of correct classification of poor or excellent data at a given cut-off score) was calculated based upon these cut-off scores (Table 3). These thresholds were subsequently used in the validation sample to determine the generalizability of these values. Accuracy of the automated QA metrics was compared using Delong method of comparing two AUCs (DeLong et al., 1988) within the proc R package (Robin et al., 2014). Z-statistics and the corresponding p-value are provided for pairwise AUC comparisons. The above analyses were performed on the full initial sample and validation sample.

The following analyses were performed on the demographically matched subsamples. Mean whole brain TSNR, FA and MD maps by QA subgroup (matched for age, sex and race) were compared using FSL's FLAME (Jenkinson et al., 2012; Smith et al., 2004; Woolrich et al., 2009). FA comparisons were restricted to major white matter tracts. In addition to whole brain analyses, mean ROI values extracted from ICBM/JHU White Matter Tract Atlas were compared across QA subgroup using MANCOVAs (Type III Sums of Squares) and follow-up t-tests (Bonferroni corrected). All statistics were performed using R (3.1.2) statistical software (R-Core-Team, 2012).

3.0 Results

3.1 Initial Sample

3.1.1 Participant Characteristics—Detailed participant characteristics of the PNC have been previously reported (Satterthwaite et al., 2014). Participant characteristics with DTI data are displayed by manual QA status (Poor, Good, Excellent) in Table 1. Overall, groups differed by age [$F(2,1352)=74.62$, $p<2.2\times10^{-16}$]. The Poor group was younger than the Good ($p=1.8\times10^{-9}$) and Excellent ($p<2.0\times10^{-16}$) groups; the Good group was younger than

[1]One individual did not complete the psychopathology assessment and was excluded from analysis.

the Excellent group ($p=1.3\times10^{-12}$). Sex differed by QA group ($\chi^2(2)=13.58$, p=0.001). There were a larger proportion of females in the Excellent group (58%) than in either the Good (49%) or Poor (46%) groups. The groups did not differ in racial distribution ($\chi^2(2)=0.75$, p=0.95) or baseline psychopathology factor score [F(2,435)=0.86, p=0.43].

3.1.2 Quality Assurance Metrics—Overall, 10.75% of the sample was manually flagged as Poor, 34.50% of data was noted as being Good and 54.75% was Excellent (Table 1). Quality assurance metrics are shown in Table 1. MEANVOX [F(2,1352)=631.77, $p<2.2\times10^{-16}$] and MAXVOX [F(2,1352)=426,75, $p<2.2\times10^{-16}$], MOTION [F(2,1352)=397.22, $p<2.2\times10^{-16}$] and TSNR [F(2,1352)=511.78, $p<2.2\times10^{-16}$] differed by QA group. Metrics were best in the Excellent group when compared to the Good group (all $ps<2\times10^{-16}$), who in turn had better metrics than the Poor group (all $ps<2\times10^{-16}$).

3.1.3 ROC analysis of automated QA measures—The ROC curve analysis was used to evaluate the diagnostic accuracy of each QA measure (MEANVOX, MAXVOX, MOTION and TSNR) to discriminate Poor and acceptable (Good + Excellent) data from each other. Graphic representations of the ROC curves are provided in Figure 4, and Table 3 details relevant cut-offs and classification accuracies for each measure. The AUC for identifying Poor data was excellent for all metrics (greater than 0.89). Overall, TSNR had the highest AUC of 0.93 (95% CI: 0.91–0.95). TSNR outperformed MAXVOX (Z=3.65, p=.0003) and MOTION (Z=2.92, p=.0036), but not MEANVOX (Z=1.11, p=0.27). The classification accuracy of TSNR was 87.03% at the optimal cut-off point of TSNR= 6.47, which was a value with 85% sensitivity, 84% specificity. Multifactor ROC analyses were explored using the R-package ‘multinom’. These analyses did not improve the AUC if additional predictors (MAXVOX, MEANVOX or MOTION) were included in the model.

The AUC and classification accuracy of these QA metrics was reasonable at identifying Good vs. Excellent data (Figure 4, Table 3). MAXVOX out-performed TSNR (Z=11.37, p<.0001), MEANVOX (5.72, p<.0001) and MOTION (7.65, p<.0001) with an AUC of 0.88. The maximum classification accuracy was 80% at a cut-off value of 2282, with 78% sensitivity and 83% specificity. Again, exploratory multifactor ROC analyses did not improve upon the AUC of MAXVOX alone.

3.2 Validation Sample

3.2.1 Participant Characteristics—Participant characteristics with DTI data at follow-up are displayed by manual QA status (Poor, Good, Excellent) in Table 2. Given the selection criteria for follow-up, there were few Poor data sets. Overall, groups differed by age [F(2,371)=7.03, p=0.001]. The Good group remained significantly younger than the Excellent group (p<.0001); post-hoc analyses indicated that Poor group was younger than the Excellent group (p<.05). The sex difference by QA group persisted in the validation sample ($\chi^2(2)=6.41$, p=0.04). There were a larger proportion of females in the Excellent group (54%) than in either the Good (38%) or Poor (50%) groups. The groups did not differ in racial distribution or baseline psychopathology score.

3.2.2 Quality Assurance Metrics—Quality assurance metrics are shown in Table 2. MEANVOX [$F(2,371)=80.77$, $p<2.2\times10^{-16}$] and MAXVOX [$F(2,371)=90.57$, $p<2.2\times10^{-16}$], MOTION [$F(2,371)=63.23$, $p<2.2\times10^{-16}$] and TSNR [$F(2,371)=93.73$, $p<2.2\times10^{-16}$] differed by QA group. Metrics were best in the Excellent group when compared to the Good group (all ps$<1.6\times10^{-8}$), which had better metrics than the Poor group (all ps<.02).

3.2.3 Validation of automated QA measure ROCs—Eighty-three percent (83%; 10/12) of Poor data and 96% (348/362) of acceptable (Good+Excellent) data was correctly identified in the validation sample using the TSNR cutoff generated in the initial sample (Table 3). Furthermore, 52% (31/60) of Good data sets and 96% (283/302) of Excellent datasets were identified using the MAXVOX intensity cut-off value (Table 3).

3.3 Data quality effects on fractional anisotropy and mean diffusivity

An age, gender, race and psychopathology matched subset (n=146, per group) was used to measure the influence of poor data quality. Characteristics of the matched sample are provided in Table 5. As in the initial sample, MEANVOX [$F(2,435)=192.66$, $p<2.2\times10^{-16}$] and MAXVOX [$F(2,435)=148.90$, $p<2.2\times10^{-16}$], MOTION [$F(2,435)=112.16$, $p<2.2\times10^{-16}$] and TSNR [$F(2,435)=284.23$, $p<2.2\times10^{-16}$] differed by QA group. In addition, the between QA group comparisons remained consistent with findings in the full sample: Excellent>Good>Poor (all ps$<6.9\times10^{-4}$).

As indicated in the ROC analyses, identification of Poor data using automated methods is reliable, especially for TSNR. TSNR maps were generated for each group in the matched sample (Figure 5). Excellent and Good data show noticeably higher TSNR throughout the cortex.

In order to determine the influence of including Poor data on group level analyses, FA and MD were compared between Poor (Fail QA) and Good+Excellent data (Pass QA). In general, data failing QA had lower FA and higher MD than data passing QA (see below); differences between the Good and Excellent data were minimal and as such no comparisons were generated.

3.3.1 TSNR—Temporal signal-to-noise ratio was significantly lower in the Poor group as compared to data passing QA (Good + Excellent) across the brain including much, but not all of white matter (Figure 6A). There were several areas within white matter where data in the Poor group showed higher TSNR (Figure 6B). Exploratory, quantitative measurement of TSNR indicated significant differences (ps<.001) in all GM ROIs, on average, and in most, but not all white matter ROIs (ps<.001). Those individuals that pass QA had higher TSNR values than those failing QA in all GM and WM regions except in the corticospinal tracts (Figure 6C&D).

3.3.2 Fractional Anisotropy—Fractional anisotropy was significantly lower throughout the major white matter tracts in the Poor group (Figure 7A). The Poor group had areas of higher FA, however these tended to be at the edges of white matter tracts or within the ventricles. This pattern suggests higher noise, possibly due to motion or scanner artifacts.

Regional, normalized FA for the Poor group is displayed in Figure 8A. FA ranged between 0.38–0.82 units below the Pass QA group (Good + Excellent).

Exploratory MANOVA analysis of ROI level data within the matched group indicated that age [$F(18, 416)=16.45$, $p<2.2\times10^{-16}$], sex [$F(18, 416)=2.88$, $p<8.03\times10^{-5}$] and race [$F(18, 416)=3.63$, $p=1.01\times10^{-6}$] each explained significant variance in FA across ROI. After accounting for these effects, FA differed between the two QA groups [$F(18, 416)=5.71$, $p<3.34\times10^{-12}$]. The Poor group consistently had lower FA (Supplemental Table 1) in all ROIs as compared to the Good + Excellent group (all ps<0.0008, Bonferonni corrected; See Supplemental Table 2). FA in the Good and Excellent group was equivalent across all ROIs. Regions of high anisotropy (e.g. corpus callosum) were negatively correlated with TSNR (Supplemental Figure 3).

3.3.3 Mean Diffusivity—Mean diffusivity was significantly higher in specific areas of the major white matter tracts in the Poor group (Figure 7B) as compared to the data passing QA. The Poor group did show areas of lower MD, however these tended to be at the edges of white matter tracts or within the ventricles. Regional, normalized MD for the Poor group is shown in Figure 8B. FA ranged between −0.13 below to 0.33 units above Pass QA group.

Since DTI data is often reported in specific white matter tracts, analyses of ROI data were also performed. ROI level data indicated that age [$F(18, 416)=14.68$, $p<2.2\times10^{-16}$] and sex [$F(18, 416)=2.56$, $p<.0005$], but not race [$F(18, 416)=1.34$, $p=0.16$] each explained significant variance in FA across ROI. After accounting for these effects, FA differed among the three QA groups [$F(18, 416)=2.76$, $p<0.0002$]. QA group differences in MD were limited to a few ROIs, including the left ($p=0.03$) and right ($p=0.008$) cingulate gyrus of the hippocampus and the right cingulum bundle ($p=0.0003$; Supplemental Table 2).

3.3.4 Age associations with FA and MD—White matter anisotropy rapidly increases with age during neonatal development and continues to increase until the middle of the third decade of life (Barnea-Goraly et al., 2005; Yoshida et al., 2013). Hence, the association between white matter FA and age was used to measure the effect of poor quality data on a known association. Across all individuals the correlation between age and average whole brain FA was 0.41 [$t(436)=9.60$, $p<2.2\times10^{-16}$]. However, the association between age and whole brain FA was lower in the Poor group ($r=0.35$), and was significantly higher in the data passing QA ($r=0.51$, $z=1.93$, $p=.05$; Figure 9A&G). A follow-up comparison across the three QA groups is shown in Supplemental Figure 4; the association between FA and age was highest in the Excellent group ($r=0.55$) and the Good group was intermediate ($r=0.47$).

Mean diffusivity typically shows the inverse pattern of anisotropy; a notable decrease until the middle of the third decade of life (Barnea-Goraly et al., 2005; Yoshida et al., 2013). This pattern is reflected in the current data in the strong negative association between FA and MD ($r=-0.71$, $t(436)=21.62$, $p<$ $p<2.2\times10^{-16}$; Figure 9B&H). The association between age and average whole brain MD was −0.48 [$t(436)=11.57$, $p<2.2\times10^{-16}$]. Again, the association between age and whole brain FA was lowest in the Poor group ($r=-0.38$), and was more robust in the group that passed QA ($r=-0.57$; $z= 2.42$, $p=.01$; Figure 9B). In a follow-up

comparison across the three QA groups the association was similar in both Good (r=−0.59) and Excellent groups (r=−0.55).

In considering whole brain and ROI data together, it is evident that a failure to rigorously quality control DTI data will result in a reduction of known associations or cause false-positive associations between typical DTI outcome measures and age.

3.4 Exploratory analysis of the role of sex on DTI artifact

An exploratory analysis of the role of sex and QA group on whole brain FA revealed a main effect of age $F(1, 433)=10.25$, $p<2.0\times10^{-16}$ (higher age with higher FA) and sex $F(1, 433)=2.92$, $p<.005$ (males>females) and an interaction of QA group by sex $F(4,433)=1.98$, $p<.05$). Follow-up comparisons of this interaction indicate that the relationship between FA and age in females passing or failing QA was no different, However, males passing QA had a higher correlation between age and FA than those failing QA ($p=.01$; Figure 9C&E). Effects for MD (Figure 9D&F) complemented these FA findings.

3.5 Comparison of QA methodology with DTIPrep

There was significant overlap between the visual QA method and DTIPrep QA (Figure 10A). Our visual QA assessment was significantly associated with DTIPrep QA outcome [$F(2,435)=48.45$, $p<2.2\times10^{-16}$]. Poor data was more likely to fail DTIPrep than Good ($p=1.2\times10^{-9}$) or Excellent ($p<2.2\times10^{-16}$) data; Good data was more likely to fail DTIPrep than Excellent ($p<.0035$) data. However, an overall comparison of binary inclusion/exclusion by visual QA group indicated significant difference in classification ($\chi^2(2) =79.80$, $p<2.2\times10^{-16}$). Concordance between QA methods (Figure 10B) was highest in the Poor (86%) data, followed by Excellent (64%) and then Good (47%) data. Surprisingly, DTIPrep excluded 44% of data we considered Good or Excellent via visual QA. This pattern is logical given that by definition more ambiguity exists in the quality of the Good data than either of the other two groups. Most data failed due to interslice brightness artifact (Figure 10C), but some data failed the interlace artifact detection (Figure 10D).

Since we earlier identified temporal SNR as a useful metric in separating high/low quality data in an automated manner, TSNR was compared across DTI passing and failing DTIPrep. As for visual QA (see Table 6), data passing DTIPrep had higher TSNR than data failing DTIPrep QA [$F(1,436)=11.03$, $p<2.2\times10^{-16}$]. Moreover, we found a significant positive correlation ($r(181)=0.32$, $p<8.2\times10^{-6}$) between TSNR and the number of volumes/gradients included in the final output from DTIPrep. In addition, MAXVOX, which was found to best differentiate Good and Excellent data, was higher in data that failed as compared to data passing DTIPrep QA, in both Good and Excellent groups (Table 6). Finally, the correlation between age and FA in data passing DTIPrep QA was lower than data passing our QA (Figure 11), but higher than data failing our QA. While this correlation was nominally lower than data passing our QA it was not significantly lower [$Z(474)= −0.82$, $p=0.41$], however, unlike our data, this correlation was not significantly different from data failing visual QA [$Z=0.96$, $p=0.34$].

4.0 Discussion

Our results highlight the confounding effects of artifacts on DTI analysis and suggest that estimates of data quality should not only be reported, but also accounted for. This is particularly true in cases where data quality is heterogeneous or is likely to be related to an outcome of interest, such as in studies of brain development or clinical psychopathology. Using DTI data from the PNC we show that specific image metrics, such as temporal signal-to-noise ratio (TSNR), can be used to reliably identify suboptimal data. Notably, automated detection not only corresponded well with manual assessment of data quality, but use of thresholds derived from the initial PNC sample were successful in identifying low-quality data in a validation sample.

4.1 The effect of including low quality data in DTI analysis

Not surprisingly, low-quality data had significantly lower fractional anisotropy throughout major white matter tracts and showed higher mean diffusivity in several brain regions as compared to data that passed quality assurance. Most importantly, the well-described positive correlation between age and FA in youth was significantly attenuated when suboptimal data was included in the overall analysis. Our findings suggest that the fidelity of DTI data can be improved by employing rigorous data quality assurance procedures, and that a failure to identify these problems may significantly impact results.

Overall, our findings in a large population-based sample indicate the importance of implementing quality control in diffusion tensor imaging analysis. Manual visualization indicated that over 10% of the data had significant artifact. While high, this is a smaller percentage than in some pediatric samples (Li et al., 2013), and is likely representative of what could be expected in adolescent samples. We show that the labor-intensive manual QA process can be automated using readily available metrics of image quality. For example, TSNR can be estimated from the data and used as a reliable measure for data exclusion. Discrimination of data considered to be of the highest quality and data with a few artifacts was more challenging, although some degree of separation was achievable. The use of TSNR can be considered a quick screen of the overall data quality, which previous studies also suggest (Chen et al., 2015; Farrell et al., 2007; Liu et al., 2015). Our results indicate that methods to evaluate TSNR are needed to consistently evaluate data fidelity within an individual protocol and, more importantly, across protocols and scanners. Our quality assurance is straightforward and uses standard tools to estimate TSNR, outliers and motion. This approach may be most useful in large samples and offers an initial approach to quickly assess the quality of a given DTI scan. Thus, quality assurance measures should be considered in DTI studies, both large and small.

In the current study, TSNR was the most discriminating QA metric, but relative motion and signal intensity outliers were nearly as effective. Yet, the lengthy acquisition time of DTI tends to reduce SNR and allow more time for motion artifact. Based on our results other measures (e.g. motion) could easily serve as substitutes with similar efficacy. In fact, the use of motion as a nuisance regressor is able to reduce the influence of spurious head motion on DTI metrics in adolescents with and without neurodevelopmental disorders (Yendiki et al., 2014). However, motion may not be the best metric, in fact our ROC analysis show that

TSNR outperformed MOTION. Often non-weighted images are frontloaded during acquisition, only a few may be acquired, and registration of data with high b-values is unreliable. Here, we used only the non-weighted images to estimate motion. This approach was feasible and robust as our non-weighted images were high quality and systematically interspersed throughout our acquisition (approximately 1 every 10 volume acquisitions). Typically SNR is computed from the b=0 images only, which alleviates the complication of varying b-values and directions (Chen et al., 2015; Farrell et al., 2007; Liu et al., 2015). However, we chose to calculate TSNR from the weighted (b>0) images. While these images are of lower SNR overall, significantly more of these images are collected, which provides a more robust estimate of SNR. In fact, TSNR from the b>0 outperformed TSNR estimated from b=0 images when classifying data that failed or passed QA (Supplement Results), although the two approaches were quite similar. Finally, the use of accelerated acquisition sequences and prospective motion correction techniques (Aksoy et al., 2011; Alhamud et al., 2012; Benner et al., 2011) will help reduce the time necessary to acquire DTI data and potentially improve motion artifact, however these sequences are also sensitive to motion and other common artifacts (Feinberg and Setsompop, 2013).

Expectedly, the inclusion of poor DTI data results in loss of data fidelity. Specifically, lower FA and higher MD were found in poor quality data compared to data passing quality assurance. Indeed, lower FA values were found throughout the large white matter tracts, whereas higher MD was more sporadic and tended to be located at tissue boundaries. These changes were robust at both the whole brain and ROI level. Most concerning was the reduction in the correlation between age and FA or MD when data that failed QA were included in the analysis. Our findings that QA may affect the relationship between age and FA, specifically in males, is also noteworthy. Young males typically show a larger association between FA and age (Wang et al., 2012), a finding we corroborate. This sex difference is thought to reflect earlier white matter maturation in females than in males. However, this relationship was smaller in males with poor quality data. Thus, our findings of an interaction with QA group and sex further underscore the importance of proper QA of DTI data. Given that the rates of white matter maturation may differ by gender, improper inclusion of data could skew the results of sex effects. Taken together, these age and sex effects provide evidence that the failure to properly QA data may contribute to erroneous associations.

This study raises an important question on implications for the experimental design of DTI studies and the power analysis of single center studies. Previous studies have addressed this issue with local rejection of data (Lauzon et al., 2013). We performed an exploratory power analysis (α =0.05; 1−β=0.80) based upon our reported relationships between age and FA in a quality-controlled sample and all data independent of quality control. Not surprisingly, significantly fewer samples are needed to estimate the age/FA correlation when using QAed (N=26) than when not (N=47; Supplemental Results & Supplemental Table 3). However, arguably the minimum sample size should also include a plan for data loss due to poor QA, particularly in samples including children and young adults.

At a minimum, one or all of these QA measures should be reported in DTI studies, especially when attempting to elucidate group difference in FA or MD. The more

appropriate approach may be to exclude data not meeting certain data quality thresholds, which is common in other neuroimaging modalities, and correcting for these factors in tensor estimations and analysis.

4.2 Comparison to DTIPrep

Other tools to investigate and correct for artifacts in DTI exist. DTIPrep and RESTORE (e.g. (Chang et al.; He et al., 2014; Li et al., 2014; Liu et al., 2015; Oguz et al., 2014) aim to investigate, eliminate and/or correct problematic slices or volumes in DTI data. Here, we directly compared our QA methods and output to that of DTIPrep. Overall, we show high concordance between our QA method and DTIPrep for Poor quality DTI scans, but we find DTIPrep to be limited by a substantial failure rate even for apparently high-quality images. While we agree that DTIPrep is a user-friendly tool that offers initial QA of DTI, several limitations of this software remain. We believe that our approach offers complementary information to DTIPrep that can be useful during quality assessment of DTI data. In the supplemental discussion we briefly discuss these points (See Supplementary Results)

Overall, our methods and those implemented in DTIPrep share the common goal of improving the quality of DTI data analyzed. DTIPrep aims to remove DTI artifact on a slicewise basis while we take a global approach to excluding potentially problematic DTI dataset in full. The latter approach is ideal for providing a quick quality assessment, particularly in large datasets and has relevant application to Big Data. We chose a global rejection threshold approach based on previous work indicating that uniform rejections resulted in fewer changes in DTI metrics (Chen et al., 2015). However, we note that it is illogical to expect one tool to robustly capture all artifact in DTI data and completely replace visual inspection, as recently indicated by Liu et al., (Liu et al., 2015). Moreover, we acknowledge the obvious limitation of globally excluding data, however, we are encouraged by the present results and believe that our method contributes significantly to the growing field of DTI QA.

4.3. Limitations

Several limitations of the current study should be noted. First, the data does not address the need for quality assurance metrics that are compatible across scanner type or diffusion protocol. While our sample is large and we include a validation, all of the current data was collected on the same scanner, using the same head coil and same DTI protocol. Thus, the generalizability of the specific cut-offs provided in the current study is limited, yet, our methodology is straightforward and easily implemented in small or large samples. While the validation sample was not selected for initial DTI quality, it was selected on the quality of each individual's structural scan—which is likely correlated with the quality of their DTI scan. Thus, there were fewer instances of Poor data in the follow-up sample. Given the lack of DTI templates that are based on samples that include adolescents, we derived a template from the current sample and then normalized white matter ROIs from an adult template to this newly derived template. While the anatomical specificity appears high, slight deviations in alignment could occur. However, this would have no influence on the whole brain findings, and we believe the normalization process was robust. Notably, only one expert manually rated each DWI volume. This was a time consuming procedure that is difficult to

implement in large-scale imaging studies such as the PNC, and further it is potentially subject to operator bias. However, it is encouraging that automated measures, both our own and as implemented in DTIPrep, largely correspond with this operator's recognition of 'bad' DWI data. The DTI protocol implemented was not specifically selected for quality assurance, as this was not the main outcome of interest in the PNC study. Recent work indicates little advantage to using a large number of encoding directions for estimating typical DTI metrics of FA and MD (Lebel et al., 2012). However, diffusion acquisitions with many gradient encoding directions have specific advantages for tract based diffusion metrics, ensuring data quality and permitting non-traditional analysis. Last, we do not address the specific contributors to increased artifact in the DTI data. As previously stated, this artifact likely has many sources including, but not limited to motion, eddy currents, thermal noise and susceptibility artifacts. As this sample contains adolescents, it is likely that head motion contributes significantly to these artifacts, especially since motion scales inversely with TSNR and individuals with poor data quality had significantly higher motion than individuals producing data that passed quality assurance. Proactive protocols that aim to reduce head-motion (e.g. incentives; (Theys et al., 2014) are likely to help improve DTI findings in samples that include adolescents.

4.4 Conclusions

Typically, DTI studies report that data with "obvious artifacts" are removed or excluded. However, a transparent, standardized estimate of quality assurance is rarely given. Few clinical or neurobiological studies report specific artifacts or quantify SNR or motion with DTI findings. This is despite prior methodological studies, which have outlined the influence of problematic DTI data on typical outcome metrics (Anderson, 2001; Armitage and Bastin, 2001; Bastin et al., 1998; Chen et al., 2015; Heim et al., 2004; Jones, 2004; Jones and Basser, 2004; Pierpaoli and Basser, 1996). These studies tend to be small and use simulated data, but confirm that artifacts in DTI data can be overcome if enough directions are collected, or if the loss of data during a given DTI acquisition is random, especially across subjects (Chen et al., 2015; Heim et al., 2004; Jones, 2004; Jones and Basser, 2004). Moreover, systematic data loss or the exclusion of data from too many diffusion directions can affect the estimates of FA, and to a lesser degree MD (Chen et al., 2015; Jones, 2004; Jones et al., 1999). Our findings of the benefits of rigorous quality assurance corroborate a recent study in a pediatric populations (Lauzon et al., 2013) and a adolescent sample including a subsample with autism (Yendiki et al., 2014). Importantly, the approach used in the current study can easily be implemented in any completed or on-going DTI study, it easily incorporated into NiFTI processing pipeline, and produce metrics that can easily be incorporated into analyses. However, most of the quality assurance strategies, including the approach reported here, rely upon post-processing correction; which comes with a loss of overall SNR, subsequently affecting measures of FA or MD (Li et al., 2013). Future approaches should consider use of DTI phantoms to identify uncertainties associated with acquisition procedures (Berberat et al., 2013) and proactive monitoring (Li et al., 2013) during DTI acquisition and informed post-processing quality assurance and analysis.

In conclusion, our results support recent findings that the use of quality assurance metrics is necessary in DTI analysis. Furthermore, we find that when quality assurance metrics are

used in the identification of suboptimal data and the data is removed, overall data fidelity increases, as does the association with age. We note the limitations in any retrospective correction of DTI artifact and do not believe that implementing our methods, or any other method, will entirely eliminate these confounds. Yet, we urge the reporting of these metrics in DTI studies and the use of improved quality assurance metrics in data analysis, especially in the face of recent Big Data initiatives.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

A special thank you to Dr. Luke Bloy for his valuable input for estimating the rotational variance of the PNC DTI encoding scheme. Thanks to the acquisition and recruitment team: Jeff Valdez, Raphael Gerraty, Marisa Riley, Jack Keefe, Elliott Yodh, R. Sean Gallagher and Rosetta Chiavacci. We thank Elena Wu-Yan for her help with data processing.

FUNDING SOURCES: This work was supported by RC2 grants from the National Institute of Mental Health MH089983 and MH089924 and P50MH096891. Additional support was provided by K01MH102609 to DRR; K23MH098130 & R01MH107703 to TDS; T32MH065218-11 to SNV, and the Dowshen Program for Neuroscience at the University of Pennsylvania. The funding sources were not directly involved in study design, collection, data analysis or interpretation, nor manuscript writing.

References

Aksoy M, Forman C, Straka M, Skare S, Holdsworth S, Hornegger J, Bammer R. Real-time optical motion correction for diffusion tensor imaging. Magnetic Resonance in Medicine. 2011; 66:366–378. [PubMed: 21432898]

Alhamud A, Tisdall MD, Hess AT, Hasan KM, Meintjes EM, van der Kouwe AJW. Volumetric navigators for real-time motion correction in diffusion tensor imaging. Magnetic resonance in medicine. 2012; 68:1097–1108. [PubMed: 22246720]

Anderson AW. Theoretical analysis of the effects of noise on diffusion tensor imaging. Magnetic Resonance in Medicine. 2001; 46:1174–1188. [PubMed: 11746585]

Armitage PA, Bastin ME. Utilizing the diffusion-to-noise ratio to optimize magnetic resonance diffusion tensor acquisition strategies for improving measurements of diffusion anisotropy. Magnetic resonance in medicine. 2001; 45:1056–1065. [PubMed: 11378884]

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011; 54:2033–2044. [PubMed: 20851191]

Barnea-Goraly N, Menon V, Eckert M, Tamm L, Bammer R, Karchemskiy A, Dant CC, Reiss AL. White matter development during childhood and adolescence: a cross-sectional diffusion tensor imaging study. Cerebral cortex. 2005; 15:1848–1854. [PubMed: 15758200]

Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophysical journal. 1994; 66:259. [PubMed: 8130344]

Basser PJ, Pajevic S. Statistical artifacts in diffusion tensor MRI (DT-MRI) caused by background noise. Magnetic Resonance in Medicine. 2000; 44:41–50. [PubMed: 10893520]

Bastin ME, Armitage PA, Marshall I. A theoretical study of the effect of experimental noise on the measurement of anisotropy in diffusion imaging. Magnetic resonance imaging. 1998; 16:773–785. [PubMed: 9811143]

Behrens TEJ, Woolrich MW, Jenkinson M, Johansen-Berg H, Nunes RG, Clare S, Matthews PM, Brady JM, Smith SM. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magnetic resonance in medicine. 2003; 50:1077–1088. [PubMed: 14587019]

Benner T, van der Kouwe AJW, Sorensen AG. Diffusion imaging with prospective motion correction and reacquisition. Magnetic resonance in medicine. 2011; 66:154–167. [PubMed: 21695721]

Berberat J, Eberle B, Rogers S, Boxheimer L, Lutters G, Merlo A, Bodis S, Remonda L. Anisotropic phantom measurements for quality assured use of diffusion tensor imaging in clinical practice. Acta Radiologica. 2013; 54:576–580. [PubMed: 23474770]

Calkins ME, Merikangas KR, Moore TM, Burstein M, Behr MA, Satterthwaite TD, Ruparel K, Wolf DH, Roalf DR, Mentch FD. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. Journal of Child Psychology and Psychiatry. 2015

Calkins ME, Moore TM, Gur RC, Satterthwaite TD, Roalf DR, Wolf DH, Merikangas KR, Burstein M, Behr JA, Ruaprel K, Mentch FD, Qui H, Chiavacci R, Connolly JJ, Sleiman PMA, Hakonarson H, Gur RE. The comorbidity of psychopathology in community youths: a symptom-level analysis from the Philadelphia Neurodevelopmental Cohort. in prep.

Calkins ME, Moore TM, Merikangas KR, Burstein M, Satterthwaite TD, Bilker WB, Ruparel K, Chiavacci R, Wolf DH, Mentch F. The psychosis spectrum in a young US community sample: findings from the Philadelphia Neurodevelopmental Cohort. World Psychiatry. 2014; 13:296–305. [PubMed: 25273303]

Chang LC, Jones DK, Pierpaoli C. RESTORE: robust estimation of tensors by outlier rejection. Magnetic Resonance in Medicine. 2005; 53:1088–1095. [PubMed: 15844157]

Chen Y, Tymofiyeva O, Hess CP, Xu D. Effects of rejecting diffusion directions on tensor-derived parameters. NeuroImage. 2015

Conturo TE, Lori NF, Cull TS, Akbudak E, Snyder AZ, Shimony JS, McKinstry RC, Burton H, Raichle ME. Tracking neuronal fiber pathways in the living human brain. Proceedings of the National Academy of Sciences. 1999; 96:10422–10427.

Conturo TE, McKinstry RC, Aronovitz JA, Neil JJ. Diffusion MRI: precision, accuracy and flow effects. NMR in Biomedicine. 1995; 8:307–332. [PubMed: 8739269]

Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical research. 1996; 29:162–173. [PubMed: 8812068]

DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988:837–845. [PubMed: 3203132]

Ding Z, Gore JC, Anderson AW. Reduction of noise in diffusion tensor images using anisotropic smoothing. Magnetic Resonance in Medicine. 2005; 53:485–490. [PubMed: 15678537]

Farrell JAD, Landman BA, Jones CK, Smith SA, Prince JL, van Zijl P, Mori S. Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging–derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 T. Journal of Magnetic Resonance Imaging. 2007; 26:756–767. [PubMed: 17729339]

Feinberg DA, Setsompop K. Ultra-fast MRI of the human brain with simultaneous multi-slice imaging. Journal of magnetic resonance. 2013; 229:90–100. [PubMed: 23473893]

Gur RC, Calkins ME, Satterthwaite TD, Ruparel K, Bilker WB, Moore TM, Savitt AP, Hakonarson H, Gur RE. Neurocognitive growth charting in psychosis spectrum youths. JAMA psychiatry. 2014; 71:366–374. [PubMed: 24499990]

Hasan KM. A framework for quality control and parameter optimization in diffusion tensor imaging: theoretical analysis and validation. Magnetic resonance imaging. 2007; 25:1196–1202. [PubMed: 17442523]

He X, Liu W, Li X, Li Q, Liu F, Rauh VA, Yin D, Bansal R, Duan Y, Kangarlu A. Automated assessment of the quality of diffusion tensor imaging data using color cast of color-encoded fractional anisotropy images. Magnetic resonance imaging. 2014; 32:446–456. [PubMed: 24637081]

Heim S, Hahn K, Sämann PG, Fahrmeir L, Auer DP. Assessing DTI data quality using bootstrap analysis. Magnetic resonance in Medicine. 2004; 52:582–589. [PubMed: 15334578]

Jack CR Jr, Bernstein MA, Borowski BJ, Gunter JL, Fox NC, Thompson PM, Schuff N, Krueger G, Killiany RJ, Decarli CS, Dale AM, Carmichael OW, Tosun D, Weiner MW. Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative. Alzheimers Dement. 2010; 6:212–220. [PubMed: 20451869]

Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage. 2002; 17:825–841. [PubMed: 12377157]

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. Fsl. NeuroImage. 2012; 62:782–790. [PubMed: 21979382]

Jones DK. The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: a Monte Carlo study†. Magnetic Resonance in Medicine. 2004; 51:807–815. [PubMed: 15065255]

Jones DK, Basser PJ. "Squashing peanuts and smashing pumpkins": How noise distorts diffusion-weighted MR data. Magnetic Resonance in Medicine. 2004; 52:979–993. [PubMed: 15508154]

Jones DK, Knösche TR, Turner R. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. Neuroimage. 2013; 73:239–254. [PubMed: 22846632]

Jones DK, Simmons A, Williams SCR, Horsfield MA. Non-invasive assessment of axonal fiber connectivity in the human brain via diffusion tensor MRI. Magnetic Resonance in Medicine. 1999; 42:37–41. [PubMed: 10398948]

Konrad K, Eickhoff SB. Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder. Hum Brain Mapp. 2010; 31:904–916. [PubMed: 20496381]

Ladouceur CD, Peper JS, Crone EA, Dahl RE. White matter development in adolescence: the influence of puberty and implications for affective disorders. Dev Cogn Neurosci. 2012; 2:36–54. [PubMed: 22247751]

Lauzon CB, Asman AJ, Esparza ML, Burns SS, Fan Q, Gao Y, Anderson AW, Davis N, Cutting LE, Landman BA. Simultaneous analysis and quality assurance for diffusion tensor imaging. PloS one. 2013; 8:e61737. [PubMed: 23637895]

Le Bihan D. Looking into the functional architecture of the brain with diffusion MRI. Nature Reviews Neuroscience. 2003; 4:469–480. [PubMed: 12778119]

Le Bihan D, Mangin JF, Poupon C, Clark CA, Pappata S, Molko N, Chabriat H. Diffusion tensor imaging: concepts and applications. Journal of magnetic resonance imaging. 2001; 13:534–546. [PubMed: 11276097]

Le Bihan D, Poupon C, Amadon A, Lethimonnier F. Artifacts and pitfalls in diffusion MRI. J Magn Reson Imaging. 2006; 24:478–488. [PubMed: 16897692]

Lebel C, Benner T, Beaulieu C. Six is enough? Comparison of diffusion parameters measured using six or more diffusion-encoding gradient directions with deterministic tractography. Magnetic Resonance in Medicine. 2012; 68:474–483. [PubMed: 22162075]

Lenroot RK, Giedd JN. Sex differences in the adolescent brain. Brain Cogn. 2010; 72:46–55. [PubMed: 19913969]

Li X, Yang J, Gao J, Luo X, Zhou Z, Hu Y, Wu EX, Wan M. A robust postprocessing workflow for datasets with motion artifacts in diffusion kurtosis imaging. PloS one. 2014; 9:e94592. [PubMed: 24727862]

Li Y, Shea SM, Lorenz CH, Jiang H, Chou MC, Mori S. Image corruption detection in diffusion tensor imaging for post-processing and real-time monitoring. PloS one. 2013; 8:e49764. [PubMed: 24204551]

Liu B, Zhu T, Zhong J. Comparison of quality control software tools for diffusion tensor imaging. Magn Reson Imaging. 2015; 33:276–285. [PubMed: 25460331]

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; 8

Oishi K, Faria AV, Yoshida S, Chang L, Mori S. Quantitative evaluation of brain development using anatomical MRI and diffusion tensor imaging. Int J Dev Neurosci. 2013; 31:512–524. [PubMed: 23796902]

Owens SF, Picchioni MM, Ettinger U, McDonald C, Walshe M, Schmechtig A, Murray RM, Rijsdijk F, Toulopoulou T. Prefrontal deviations in function but not volume are putative endophenotypes for schizophrenia. Brain. 2012; 135:2231–2244. [PubMed: 22693145]

Pierpaoli C, Basser PJ. Toward a quantitative assessment of diffusion anisotropy. Magnetic resonance in Medicine. 1996; 36:893–906. [PubMed: 8946355]

Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage. 2012; 59:2142–2154. [PubMed: 22019881]

R-Core-Team. R: A Language and Enviroment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2012.

Radanovic M, Pereira FR, Stella F, Aprahamian I, Ferreira LK, Forlenza OV, Busatto GF. White matter abnormalities associated with Alzheimer's disease and mild cognitive impairment: a critical review of MRI studies. Expert Rev Neurother. 2013; 13:483–493. [PubMed: 23621306]

Reese TG, Heid O, Weisskoff RM, Wedeen VJ. Reduction of eddy-current-induced distortion in diffusion MRI using a twice-refocused spin echo. Magnetic Resonance in Medicine. 2003; 49:177–182. [PubMed: 12509835]

Reuter M, Tisdall MD, Qureshi A, Buckner RL, van der Kouwe AJW, Fischl B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. NeuroImage. 2015; 107:107–115. [PubMed: 25498430]

Roalf DR, Gur RE, Verma R, Parker WA, Quarmley M, Ruparel K, Gur RC. White matter microstructure in schizophrenia: Associations to neurocognition and clinical symptomatology. Schizophrenia research. 2015; 161:42–49. [PubMed: 25445621]

Robin X, Turck N, Hainard A, Tiberti N, Lisacke F, Sanchez J, Muller M. display and analyaze ROC curves ('pROC'). 2014

Satterthwaite TD, Elliott MA, Ruparel K, Loughead J, Prabhakaran K, Calkins ME, Hopson R, Jackson C, Keefe J, Riley M. Neuroimaging of the Philadelphia neurodevelopmental cohort. Neuroimage. 2014; 86:544–553. [PubMed: 23921101]

Satterthwaite TD, Wolf DH, Loughead J, Ruparel K, Elliott MA, Hakonarson H, Gur RC, Gur RE. Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. Neuroimage. 2012; 60:623–632. [PubMed: 22233733]

Skare S, Hedehus M, Moseley ME, Li TQ. Condition number as a measure of noise performance of diffusion tensor data acquisition schemes with MRI. Journal of Magnetic Resonance. 2000; 147:340–352. [PubMed: 11097823]

Smith SM. Fast robust automated brain extraction. Human Brain Mapping. 2002; 17:143–155. [PubMed: 12391568]

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 2004; 23:S208–S219. [PubMed: 15501092]

Theys C, Wouters J, Ghesquiere P. Diffusion Tensor Imaging and Resting-State Functional MRI-Scanning in 5-and 6-Year-Old Children: Training Protocol and Motion Assessment. PloS one. 2014; 9:e94019. [PubMed: 24718364]

Tournier JD, Mori S, Leemans A. Diffusion tensor imaging and beyond. Magnetic Resonance in Medicine. 2011; 65:1532–1556. [PubMed: 21469191]

Travers BG, Adluru N, Ennis C, Tromp do PM, Destiche D, Doran S, Bigler ED, Lange N, Lainhart JE, Alexander AL. Diffusion tensor imaging in autism spectrum disorder: a review. Autism Res. 2012; 5:289–313. [PubMed: 22786754]

Van Dijk KRA, Sabuncu MR, Buckner RL. The influence of head motion on intrinsic functional connectivity MRI. Neuroimage. 2012; 59:431–438. [PubMed: 21810475]

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TE, Bucholz R, Chang A, Chen L, Corbetta M, Curtiss SW, Della Penna S, Feinberg D, Glasser MF, Harel N, Heath AC, Larson-Prior L, Marcus D, Michalareas G, Moeller S, Oostenveld R, Petersen SE, Prior F, Schlaggar BL, Smith SM, Snyder AZ, Xu J, Yacoub E. The Human Connectome Project: a data acquisition perspective. Neuroimage. 2012; 62:2222–2231. [PubMed: 22366334]

Wang Y, Adamson C, Yuan W, Altaye M, Rajagopal A, Byars AW, Holland SK. Sex differences in white matter development during adolescence: a DTI study. Brain research. 2012; 1478:1–15. [PubMed: 22954903]

Wheeler AL, Voineskos AN. A review of structural neuroimaging in schizophrenia: from connectivity to connectomics. Frontiers in Human Neuroscience. 2014; 8:653. [PubMed: 25202257]

Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, Beckmann C, Jenkinson M, Smith SM. Bayesian analysis of neuroimaging data in FSL. Neuroimage. 2009; 45:S173–S186. [PubMed: 19059349]

Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage. 2001; 14:1370–1386. [PubMed: 11707093]

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

Yoshida S, Oishi K, Faria AV, Mori S. Diffusion tensor imaging of normal brain development. Pediatr Radiol. 2013; 43:15–27. [PubMed: 23288475]

Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3:32–35. [PubMed: 15405679]

Zhang B, Xu Y, Zhu B, Kantarci K. The role of diffusion tensor imaging in detecting microstructural changes in prodromal Alzheimer's disease. CNS Neurosci Ther. 2014; 20:3–9. [PubMed: 24330534]

Zhang H, Yushkevich PA, Alexander DC, Gee JC. Deformable registration of diffusion tensor MR images with explicit orientation optimization. Medical image analysis. 2006; 10:764–785. [PubMed: 16899392]

Highlights

- Derives metrics of data quality assurance in a developmental sample of over 1500 DTI scans
- Temporal signal to noise ratio reliably differentiated high quality and low quality data.
- Failure to remove low quality data impacts typical DTI measures (FA/MD).
- Developmental effects on FA/MD are reduced when impact of data quality is not considered.

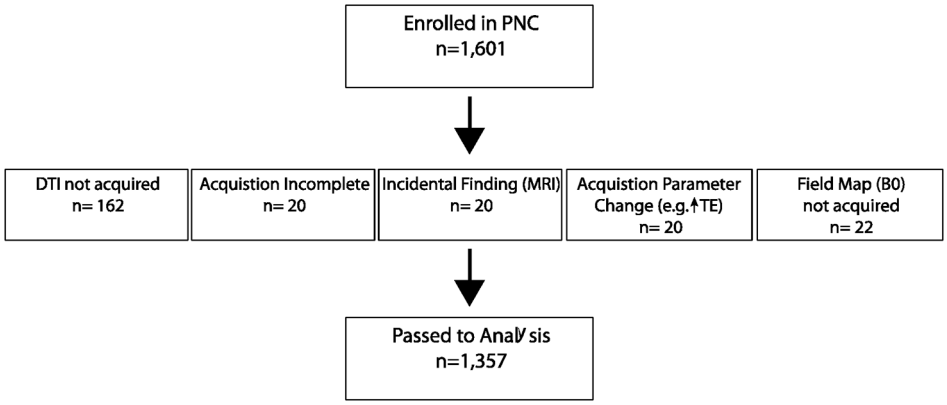

Figure 1.

Initial enrollment in the Philadelphia Neurodevelopmental Cohort and a count and explanation of excluded DTI data.

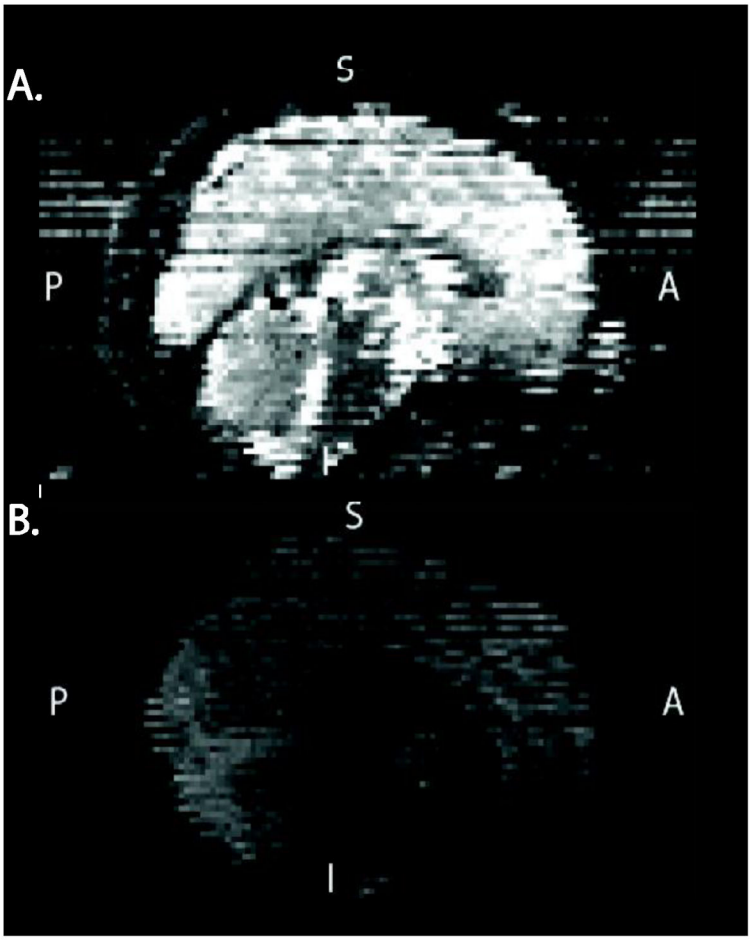

Figure 2.
Two examples of Poor DTI data from the PNC. A. An example of image striping likely cause by sub-optimal gradient performance. B. An example of data with inter-slice and intra-slice signal drop-out likely caused by the interaction of subject motion and diffusion encoding.

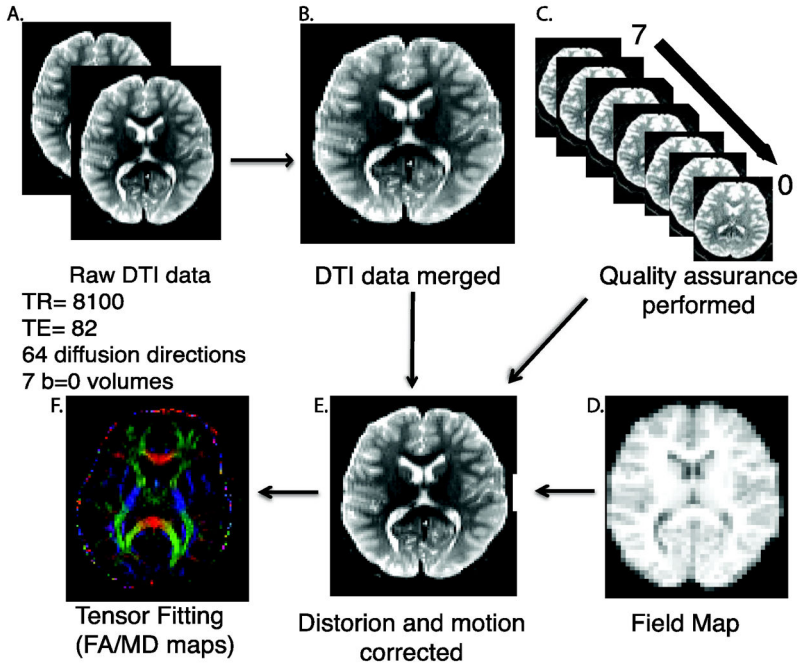

Figure 3.

DTI Preprocessing Pipeline. 64 direction DTI data was collected in two consecutive series (A) and merged into a single time series (B). Automated quality assurance was performed (C). A field map (D) was acquired and used in distortion correction. DTI images were corrected for motion and eddy currents (E). Last, a tensor model was fit (F).

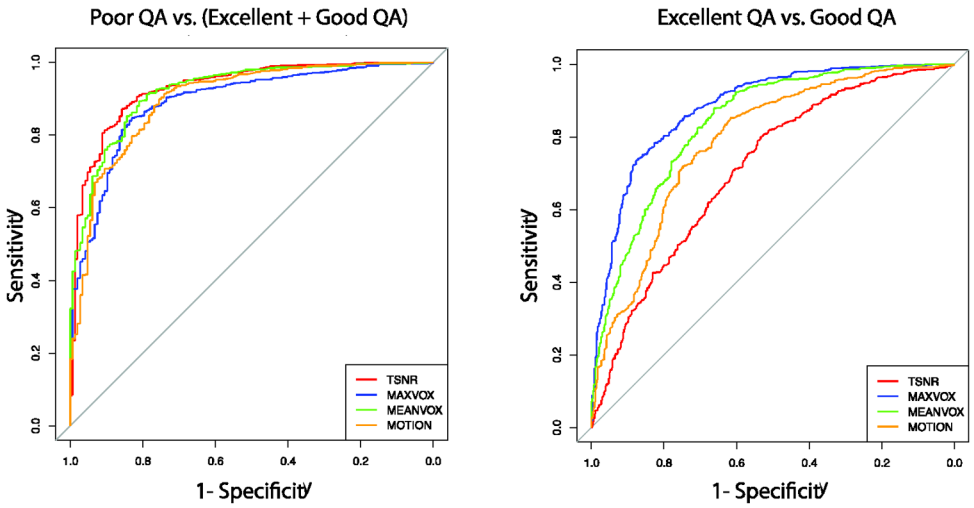

Figure 4.

Receiver operator characteristic curves of TSNR, MAXVOX, MEANVOX, and MOTION. (A) ROCs for differentiating Poor data from Acceptable data (Good+Excellent). TSNR best differentiated Poor data from Acceptable data. (B) ROCs for differentiating Excellent data from Good data. MAXVOX best differentiated Excellent data from Good Data.

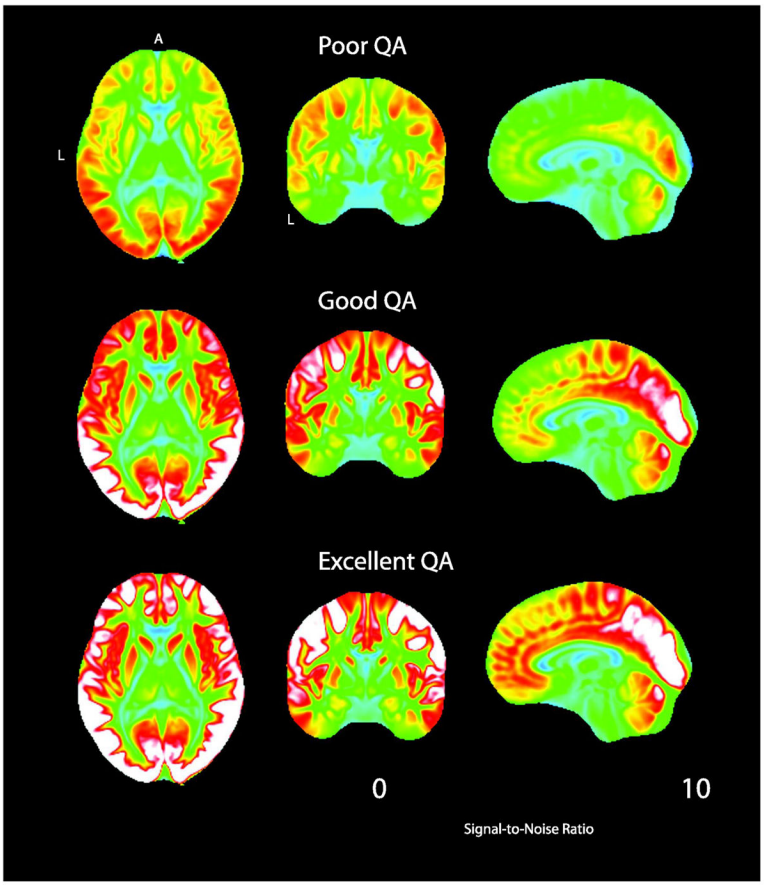

Figure 5.
Temporal signal-to-noise ratio (TSNR) maps for each QA group (matched sub-sample: n=146 per group).

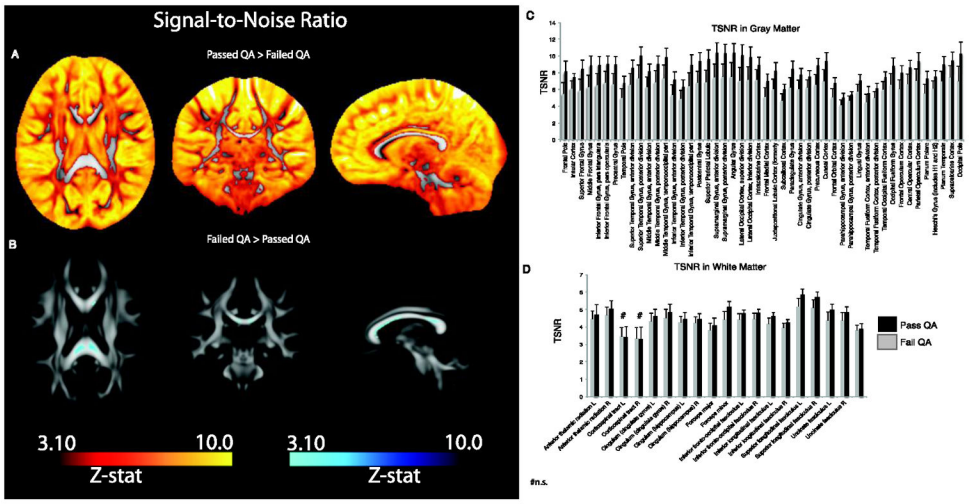

Figure 6.

Statistical differences (z-maps) in TSNR between data passing QA (Good+Excellent) in comparison to those failing QA (Poor data). A. Z-maps indicating higher TSNR across the brain in data passing QA as compared to data failing QA B. Data failing QA shows limited regions where TSNR is higher than data passing QA. This difference is limited to the corpus callosum, a highly anisotropic region, where TSNR tends to be low in all individuals. C. Quantified regional TSNR in gray matter ROIs (Harvard-Oxford atlas) in data passing and failing QA. Data passing QA had significantly higher TSNR in all ROIs. D. Quantified regional TSNR in white matter ROIs (JHU atlas) in data passing and failing QA. Data passing QA had significantly higher TSNR in all ROIs except the corticospinal tract.

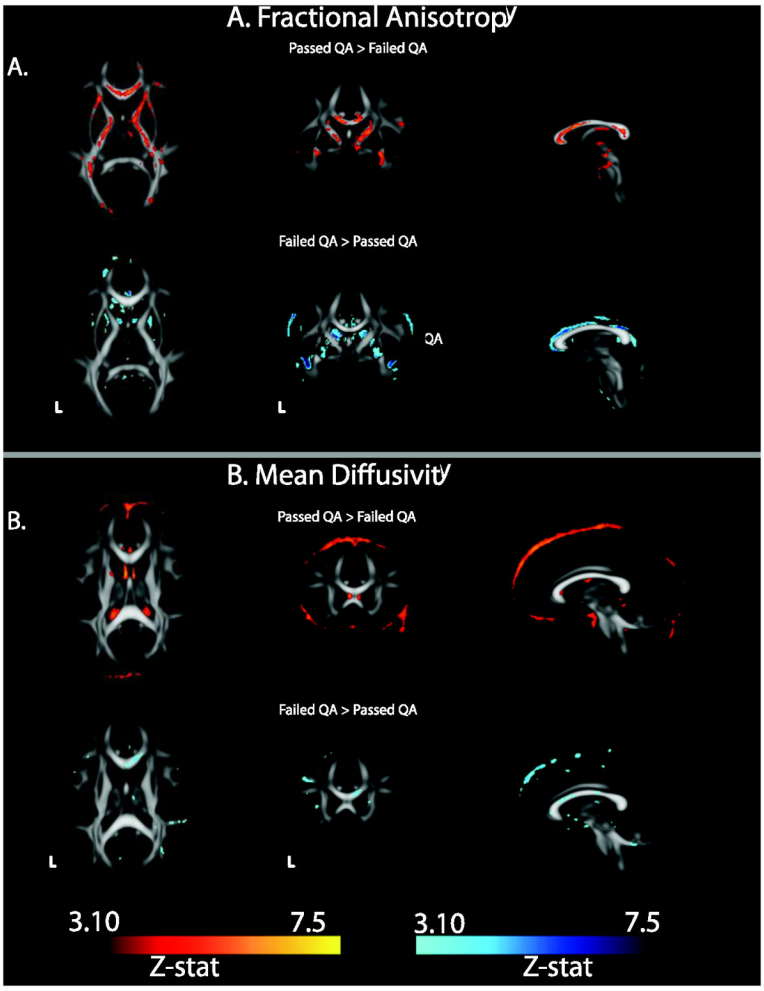

Figure 7.

Statistical difference in FA and MD between data passing QA (Good+Excellent) in comparison to those failing QA (Poor data). (A) FA was significantly higher throughout white matter in higher quality data while (B) MD was higher in Poor data, particularly at the edge of white matter tracts.

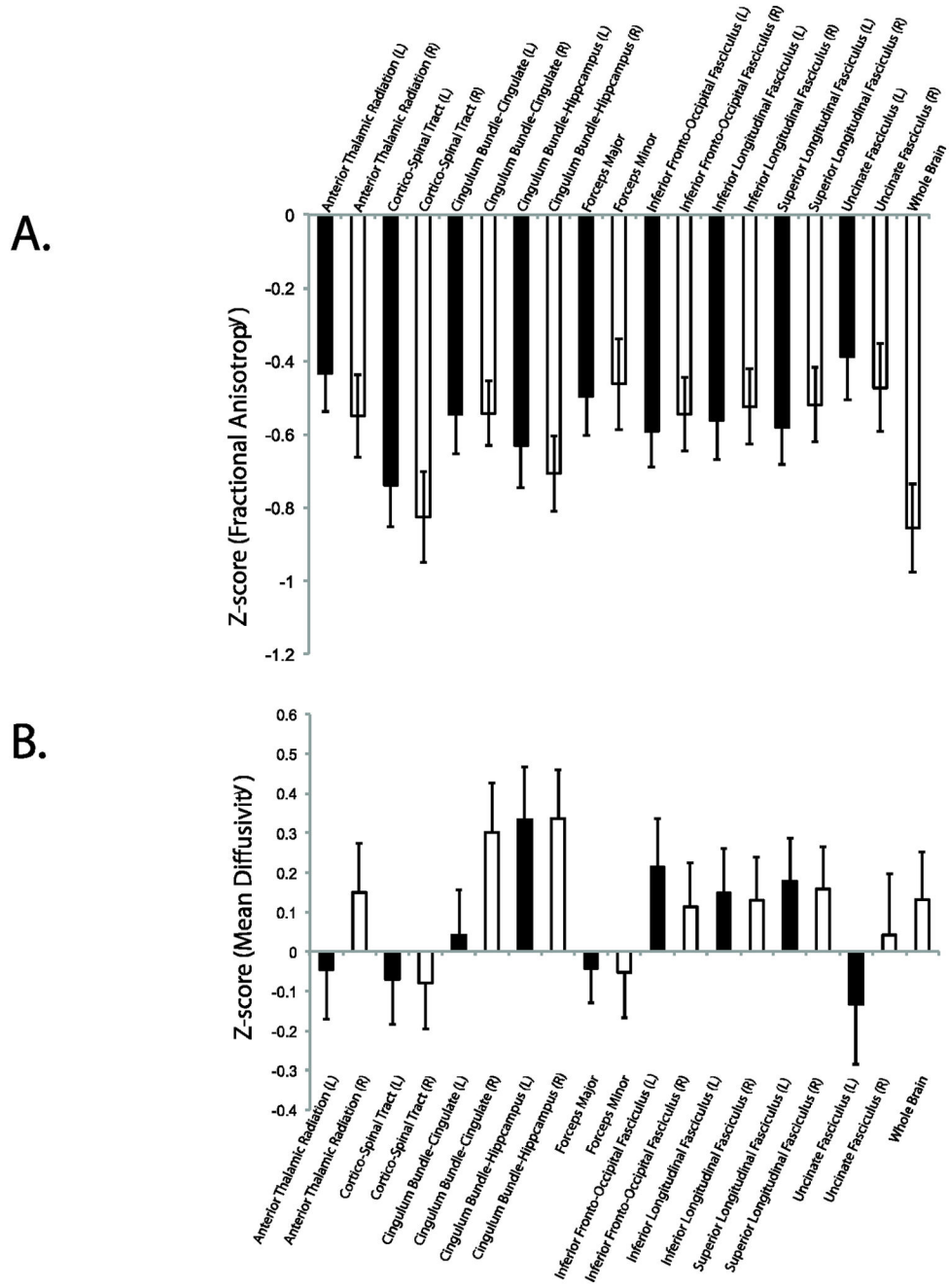

Figure 8.

Normalized (z-transformed) regional (A) FA and (B) MD values in Poor data. ROI data from ICBM-JHU White Matter Atlas. Data are normalized data that passes QA.

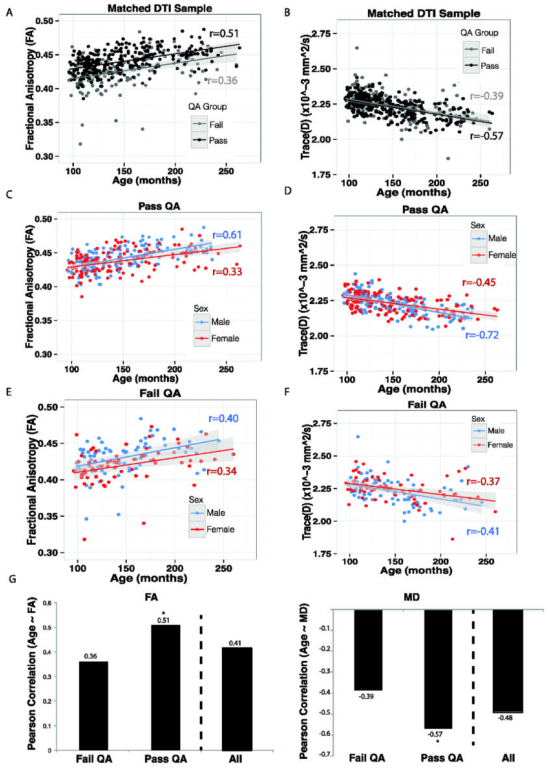

Figure 9.

Pearson correlation coefficients between Age and FA in data failing QA (Poor) and data passing QA (Good+Excellent). A. Data that failed QA had a significantly lower correlation between age and FA as compared to data that passed QA. B. Data that failed QA had a significantly lower correlation between age and MD as compared to data that passed QA. C & D: Correlation between FA and age in males and females in data passing QA. E & F: Correlation between FA and age in males and females in data failing QA. G & H. Summary plot of the correlation between age and DTI metric in data failing QA, passing QA and all data combined. The striped bar shows Pearson correlation when all data is combined. Estimating this association is strongest when data of low quality is excluded from analysis. *p<.05 as compared to Failed QA or All.

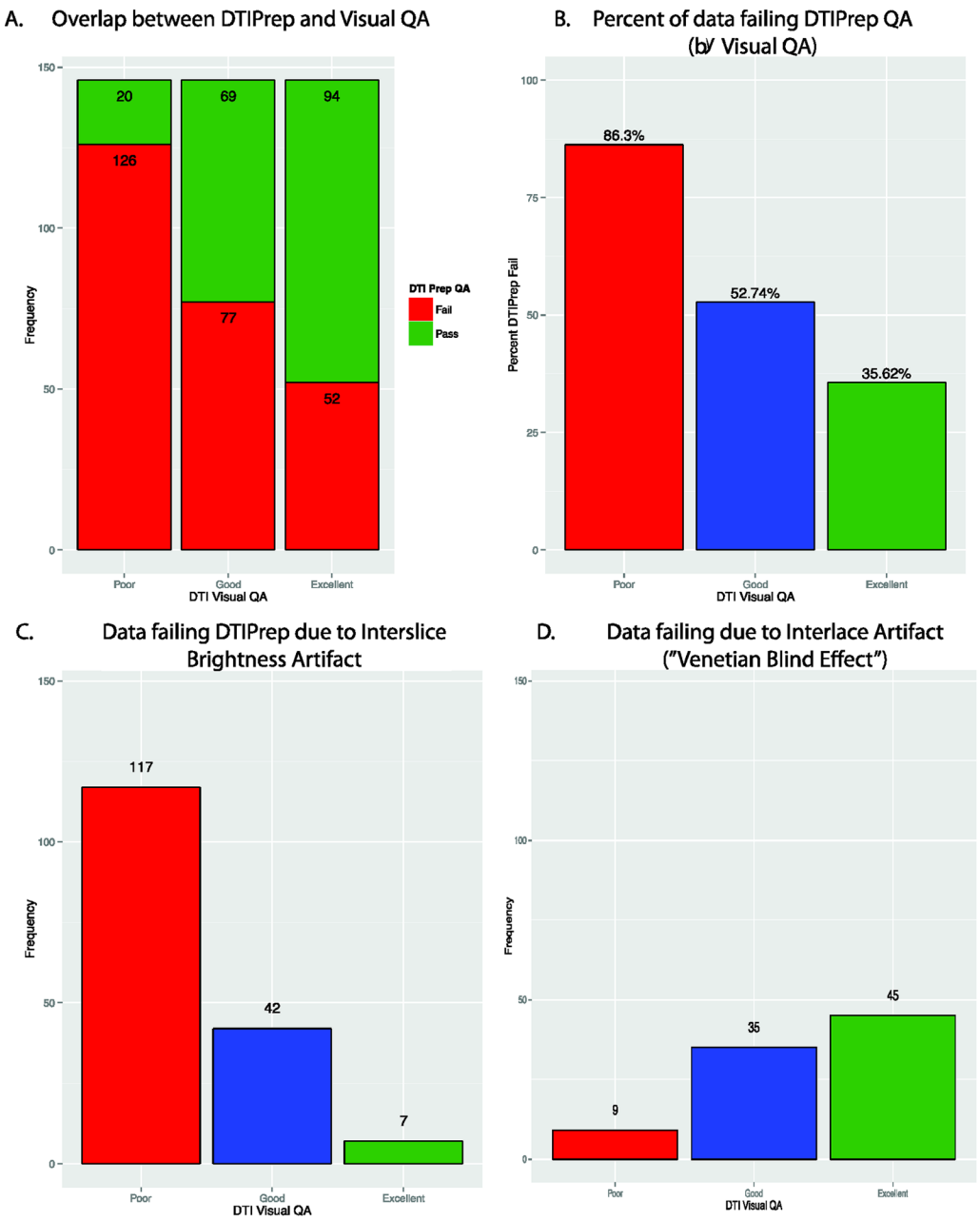

Figure 10.

Comparison of DTIPrep QA with our Visual QA methods. The two methods were in the highest agreement for Poor data followed by Good then Excellent. However, DTIPrep at the default setting considered more data to fail.

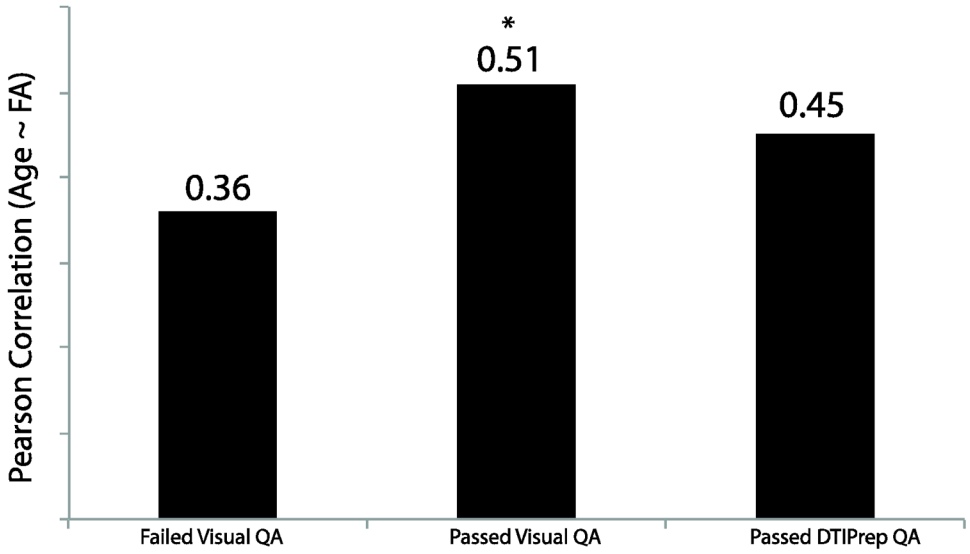

Figure 11.

Pearson correlation coefficients between Age and FA in data failing visual QA (Poor), data passing visual QA (Good+Excellent) and data Passing DTIPrep QA. Data passing visual QA, but not DTIPrep QA, had a significantly higher correlation between Age and FA. *p<0.05 as compared to Failed Visual QA.

Table 1

Initial Sample Characteristics

	Poor	Good	Excellent
	147	468	742
Sex, n			
Male	80	239	312
Female	67*	229*	430
Race, n			
Caucasian	68	206	346
African American	62	207	312
Other	17	55	84
Age, mean (SD) in years	12.7 (3.50)*#	14.56 (3.52)*	16.07 (3.25)
Baseline Psychopathology Factor Score	−0.05 (0.63)	0.01 (0.55)	0.13 (0.54)
TSNR, mean (SD)	5.52 (0.93)*	6.9 (0.68)*	7.37 (0.55)
MAXVOX, mean (SD)	14497 (8667)*#	7165 (7189)*	1684 (1741)
MEANVOX, mean (SD)	2001.50 (1080.20)*#	830.40 (597.10)*	378 (164.10)
MOTION, mean (SD)	1.89 (1.53)*#	0.68 (0.48)*	0.34 (0.28)

*p<.05 as compared to the Excellent group;

#p<.05 as compared to the Good group.

Table 2

Follow-Up Sample Characteristics

	Poor	Good	Excellent
	12	60	302
Sex, n			
Male	8	37	139
Female	4*	23*	163
Race, n			
Caucasian	6	23	134
African American	4	30	133
Other	2	7	35
Age, mean (SD) in years	15.77 (3.66) *#	15.48 (3.95)*	17.14 (3.12)
Baseline Psychopathology Factor Score	0.07 (0.47)	−0.17 (0.64)	0.03 (0.57)
TSNR, mean (SD)	4.88 (1.09) *#	6.46 (0.79)*	7.12 (0.03)
MAXVOX, mean (SD)	13037 (11676) *#	9398 (6670)*	2632 (2949)
MEANVOX, mean (SD)	2048.40 (1559.50) *#	1084.80 (952.90)*	459.70 (279)
MOTION, mean (SD)	1.9 (1.42) *#	0.84 (0.61)*	0.41 (0.41)

*p<.05 as compared to the Excellent group;

#p<.05 as compared to the Good group.

Table 3

A. ROC Measures (Initial Sample)		Poor vs (Good + Excellent)	Good vs Excellent
MAXVOX	AUC (+− 95% CI)	0.892 (0.866–0.919)	**0.881 (0.862–0.901)**##
	Sensitivity/Specificity	0.848/0.830	**0.782/0.827**
	Youden Index	0.678	**0.609**
	Cutoff	7041	**2282**
	Classification Accuracy	84.60%	**79.92%**
MEANVOX	AUC (+− 95% CI)	0.922 (0.899–0.944)	0.839 (0.816–0.862)
	Sensitivity/Specificity	0.852/0.844	0.790/0.733
	Youden Index	0.696	0.523
	Cutoff	866.9	426.1
	Classification Accuracy	85.11%	76.78%
MOTION	AUC (+− 95% CI)	0.899 (0.871–0.927)	0.787 (0.761–0.814)
	Sensitivity/Specificity	0.881/0.769	0.720/0.748
	Youden Index	0.65	0.468
	Cutoff	0.824	0.411
	Classification Accuracy	86.66%	73.22%
TSNR	AUC (+− 95% CI)	**0.93 (0.908–0.952)****	0.711 (0.68–0.741)
	Sensitivity/Specificity	**0.873/0.857**	0.706/0.611
	Youden Index	**0.73**	0.317
	Cutoff	**6.47**	7.16
	Classification Accuracy	**87.03%**	66.86%

B. Classification Accuracy of QA Measures				
		Poor	Good	Excellent
Initial Sample	Stage 1	126 (86%)	1055 (87%)	
	Stage 2	-	385 (82%)	581 (78%)
Follow-up Sample	Stage 1	10 (83%)	348 (96%)	
	Stage 2	-	31 (52%)	283 (94%)

** Significantly higher AUC than MAXVOX and MOTION

Significantly higher AUC than TSNR, MEANVOX, and MOTION

Table 4

DTI encoding scheme for the Philadelphia Neurodevelopmental Cohort.

Set #1: 32 directions + 3 b=0				Set #2: 32 directions + 4 b=0			
Vector	**X**	**Y**	**Z**	**Vector**	**X**	**Y**	**Z**
1	0.000	0.000	0.000	1	0.000	0.000	0.000
2	1.000	0.000	0.000	2	0.949	−0.233	0.211
3	0.000	1.000	0.000	3	0.267	0.960	−0.085
4	−0.026	0.649	0.760	4	−0.103	0.822	0.560
5	0.591	−0.766	0.252	5	0.774	−0.604	0.190
6	−0.236	−0.524	0.818	6	−0.142	−0.725	0.674
7	−0.893	−0.259	0.368	7	−0.827	−0.521	0.213
8	0.796	0.129	0.591	8	0.709	0.408	0.575
9	0.234	0.930	0.284	9	0.514	0.840	0.174
10	0.936	0.140	0.324	10	0.888	0.417	0.193
11	0.506	−0.845	−0.175	11	0.453	−0.889	0.068
12	0.000	0.000	0.000	12	0.000	0.000	0.000
13	0.346	−0.848	−0.402	13	0.116	−0.963	−0.245
14	0.457	−0.631	−0.627	14	0.290	−0.541	−0.789
15	−0.487	−0.389	0.782	15	−0.713	−0.247	0.656
16	−0.618	0.673	0.407	16	−0.740	0.388	0.549
17	−0.577	−0.105	−0.810	17	−0.306	−0.199	−0.931
18	0.894	−0.040	−0.447	18	0.838	−0.458	−0.296
19	−0.800	0.403	−0.444	19	−0.789	0.153	−0.596
20	0.233	0.783	0.577	20	0.502	0.690	0.521
21	−0.021	−0.188	−0.982	21	0.001	0.077	−0.997
22	0.217	−0.956	0.199	22	0.037	−0.902	0.430
23	0.000	0.000	0.000	23	0.000	0.000	0.000
24	−0.161	0.356	0.921	24	−0.381	0.143	0.914
25	−0.147	0.731	−0.666	25	−0.184	0.392	−0.901
26	−0.562	0.232	−0.794	26	−0.282	0.145	−0.948

Set #1: 32 directions + 3 b=0				Set #2: 32 directions + 4 b=0			
Vector	**X**	**Y**	**Z**	**Vector**	**X**	**Y**	**Z**
27	−0.332	−0.130	0.934	27	−0.088	−0.335	0.938
28	−0.963	−0.265	0.044	28	−0.721	−0.693	0.009
29	−0.960	0.205	0.193	29	−0.773	0.628	0.088
30	−0.693	0.024	0.721	30	−0.552	−0.792	0.259
31	0.682	0.529	−0.506	31	0.433	0.682	−0.589
32	0.584	−0.596	0.551	32	0.363	−0.561	0.744
33	−0.171	−0.509	−0.844	33	0.570	−0.303	−0.763
34	0.463	0.423	0.779	34	0.721	0.608	0.332
35	0.385	−0.809	0.444	35	0.260	0.885	−0.387
-	-	-	-	36	0.000	0.000	0.000

Table 5

Matched Sample Characteristics

	Poor	Good	Excellent
	146	146	146
Sex, n			
Male	79	78	75
Female	67	68	71
Race, n			
Caucasian	68	69	70
African American	62	62	60
Other	16	15	16
Age, mean (SD) in years	12.31 (3.26)	12.37 (3.24)	12.42 (3.21)
Psychopathology Factor Score, mean (SD)	−0.05 (0.63)	−0.10 (0.56)	−0.01 (0.58)
TSNR, mean (SD)	5.52 (0.1693)*#	6.96 (0.71)*	7.56 (0.57)
MAXVOX, mean (SD)	14479.61 (8693.96) *#	7565.69 (6504.56)*	1687.30 (1647.01)
MEANVOX, mean (SD)	2000.34 (1083.82) *#	854.76 (617.47)*	379.91
MOTION, mean (SD)	1.89 (1.53) *#	0.72 (0.50)*	0.31 (0.19)

Table 6

Visual QA status (n=438)	DTIPrep QA Status	
	Fail	Pass
TSNR		
Poor	5.38 (0.91)	6.40 (0.53)
Good	6.79 (0.73)	7.15 (0.63)
Excellent	7.46 (0.57)	7.62 (0.56)
MAXVOX		
Poor	15681 (7725)	6909 (10688)
Good	9258 (7595)	5677 (4343)
Excellent	1751 (1522)	1651 (1719)
MEANVOX		
Poor	2183 (1014)	848 (762)
Good	1039 (735)	648 (356)
Excellent	400 (180)	368 (124)
MOTION		
Poor	2.08 (1.55)	0.70 (0.54)
Good	0.86 (0.59)	0.55 (0.32)
Excellent	0.36 (0.22)	0.28 (0.16)