MASiVar: Multisite, Multiscanner, and Multisubject Acquisitions for Studying Variability in Diffusion Weighted Magnetic Resonance Imaging

Leon Y. Cai[1], Qi Yang[2], Praitayini Kanakaraj[2], Vishwesh Nath[2], Allen T. Newton[3,4], Heidi A. Edmonson[5], Jeffrey Luci[6,7], Benjamin N. Conrad[8,9], Gavin R. Price[9], Colin B. Hansen[2], Cailey I. Kerley[2], Karthik Ramadass[2], Fang-Cheng Yeh[10], Hakmook Kang[11], Eleftherios Garyfallidis[12], Maxime Descoteaux[13], Francois Rheault[2,13], Kurt G. Schilling[3,4], and Bennett A. Landman[1,2,3,4]

[1] Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, USA

[2] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

[3] Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, TN, USA

[4] Vanderbilt University Institute of Imaging Science, Vanderbilt University, Nashville, TN, USA

[5] Department of Radiology, Mayo Clinic, Rochester, MN, USA

[6] Department of Biomedical Engineering, The University of Texas at Austin, TX, USA

[7] Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, New Jersey, USA

[8] Neuroscience Graduate Program, Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, TN, USA

[9] Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, TN, USA

[10] Department of Neurological Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[11] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[12] Department of Intelligent Systems Engineering, Indiana University, Bloomington, IN, USA

[13] Department of Computer Science, Université de Sherbrooke, Sherbrooke, Québec, Canada

Keywords: Diffusion MRI, variability, reproducibility, tensor, microstructure, DTI, NODDI, bundle segmentation, connectome

Word Count: 4995

Correspondence: Leon Y. Cai
Medical-image Analysis and Statistical Interpretation (MASI) Lab
Vanderbilt University
leon.y.cai@vanderbilt.edu

ABSTRACT

Purpose

Diffusion weighted imaging (DWI) allows investigators to identify microstructural differences between subjects, but variability due to session and scanner biases is still a challenge.

Methods

To investigate DWI variability, we present MASiVar, a multisite dataset consisting of 319 diffusion scans acquired at 3T from b = 1000 to 3000 s/mm^2 across 14 healthy adults, 83 healthy children (5 to 8 years), three sites, and four scanners as a publicly available, preprocessed, and de-identified dataset. With the adult data, we demonstrate the capacity of MASiVar to simultaneously quantify the intrasession, intersession, interscanner, and intersubject variability of four common DWI processing approaches: (1) a tensor signal representation, (2) a multi-compartment neurite orientation dispersion and density model, (3) white matter bundle segmentation, and (4) structural connectomics. Respectively, we evaluate region-wise fractional anisotropy (FA), mean diffusivity, and principal eigenvector; region-wise cerebral spinal fluid volume fraction, intracellular volume fraction, and orientation dispersion index; bundle-wise shape, volume, length and FA; and connectome correlation and maximized modularity, global efficiency, and characteristic path length.

Results

We plot the variability in these measures at each level and find that variability generally increases with intrasession to intersession to interscanner to intersubject effects across all processing approaches and that sometimes interscanner variability can approach intersubject variability.

Conclusions

This study demonstrates the potential of MASiVar to investigate DWI variability across multiple levels and processing approaches simultaneously and suggests harmonization between scanners for multisite analyses is critical prior to inference of group differences on subjects.

INTRODUCTION

Diffusion weighted MRI imaging (DWI) is a noninvasive way of elucidating the brain's microstructural makeup (1). Common modes of DWI analysis include representing the diffusion signal with tensors (2,3), representing biological tissues with multi-compartment models (4–6), identifying white matter bundles (7), and investigating the human structural connectome (8). These approaches form the basis for many studies including those investigating a wide range of neurological disorders including autism (9,10), diabetes (11,12), multiple sclerosis (13), and schizophrenia (14) as well as differences due to aging (15) and sex (16). These types of studies, however, rely on the identification of group differences with respect to an independent variable. Often this variable reflects whether the scanned subject has a particular disease, or the age or sex of the subject. Robust study design can control for additional subject-level confounders through age- and sex-matching and related approaches. However, one level of potential confounding in DWI studies that has not been thoroughly characterized is the variability of calculations due to differences within and between imaging sessions and scanners.

One particular reason for this is the difficulty in acquiring data configured to perform such a characterization. For instance, to quantify variation within a session, imaging sessions with repeated scans are needed. To quantify variation between sessions and between scanners, multiple imaging sessions on at least one scanner and at least one imaging session on multiple scanners are required, respectively. Last, to assess the session and scanner effects relative to the subject effect size, multiple scanned subjects are needed as well.

Another reason for this is the low number of properly configured publicly available datasets. Some of the few that exist that allow for investigations of DWI variability are the MASSIVE dataset (17), the Human Connectome Project (HCP) 3T dataset (18), the MICRA dataset (19), the SIMON dataset (20), and the multisite dataset published by Tong et al. (21). MASSIVE consists of one subject scanned repeatedly on one scanner (17); HCP consists of multiple subjects with multiple acquisitions per session all on one scanner (18); MICRA consists of multiple subjects scanned repeatedly on one scanner (19); SIMON consists of one subject scanned at over 70 sites (20), and the Tong dataset consists of multiple subjects each scanned on multiple scanners (21).

These difficulties have resulted in existing DWI variability studies that are largely limited in scope and that offer a fragmented view of the variability landscape (Table 1). Many of these studies each capture portions of the spectrum of effects due to session, scanner, and subject biases, but are unable to assess for all levels at once. In addition, most of the existing investigations each focus on one specific DWI processing approach and/or model and as such do not provide a holistic assessment of DWI variability. As such, the understanding of how one study's variability estimates in tensor-based metrics between sessions might compare to another's estimates of tractography biases

between scanners is not obvious, for instance. Thus, to bring the field toward a more global understanding of DWI variability, the release of additional publicly available datasets configured to characterize DWI variability and a global analysis of variability on multiple levels and across different processing approaches is needed.

To fill the first need, we propose MASiVar, a multisite, multiscanner, and multisubject dataset able to characterize DWI variability due to session, scanner, and subject effects. To fill the second need, we demonstrate the potential of MASiVar to characterize DWI variability by presenting a simultaneous quantification and comparison of these effects on four different common diffusion approaches, hypothesizing that variability increases with session, scanner, and subject effects.

METHODS

Data acquisition

MASiVar consists of data acquired from 2016 to 2020 to study both DWI variability and other phenomena. As such, the data exist in four cohorts, designated I, II, III, and IV (Figure 1).

Cohort I consists of one healthy adult subject (male, age 25 years) with multiple imaging sessions on a 3T Philips Achieva scanner at site 1 (scanner A). This subject underwent three imaging sessions, one each consecutive day, and received two to three scans during each session (Figure 1). Each scan consisted of 96-direction acquisitions at b = 1000, 1500, 2000, 2500, and 3000 s/mm^2 (Table 2). These scans were acquired at 2.5mm isotropic resolution with an echo time (TE) and repetition time (TR) of TE / TR = 94ms / 2650ms.

Cohort II consists of five healthy adult subjects (3 male, 2 female, age 27 to 47 years) scanned for one to two sessions on each of three to four different scanners. Each subject underwent all sessions within one year. The scanners included scanner A, another 3T Philips Achieva scanner at site 1 (scanner B), a 3T General Electric Discovery MR750 scanner at site 2, and a 3T Siemens Skyra scanner at site 3 (Figure 1). For each imaging session, each subject received one scan, consisting of 96-direction acquisitions at b = 1000, 1500, 2000, 2500 (or 2465 at site 3 due to hardware limitations) s/mm^2 and a 30- or 32-direction acquisition at b = 1000 s/mm^2 (Table 2). The scans acquired on scanner B, at site 2, and at site 3, and all the 30- or 32-direction scans were acquired at 2.5mm isotropic resolution. On scanner A, one subject's 96-direction acquisitions were also acquired at 2.5mm isotropic resolution while the remainder were acquired at 1.9mm by 1.9mm by 2.2mm (sagittal, coronal, and axial) resolution. For acquisitions on scanner A, the 2.5mm isotropic 96-direction scans were acquired with TE / TR = 90ms / 5200ms, while the other 96-direction acquisitions were acquired with TE / TR = 90ms / 5950ms, and TE / TR = 55ms /

6127ms to 7309ms for the 32-direction acquisitions. For acquisitions on scanner B, the 96-direction scans were acquired with TE / TR = 90ms / 5800ms or 5900ms, while the 32-direction acquisitions were acquired with TE / TR = 55ms / 7022ms to 7069ms. For the 96-direction acquisitions acquired at site 2, TE / TR = 90ms / 5800ms or 5900ms, while the 32-direction acquisitions were acquired with a TE / TR of either 58ms / 7042ms or 59ms / 4286ms. All scans acquired at site 3 were acquired with TE / TR = 95ms / 6350ms. All sessions acquired on scanner A that contained scans of varying resolution were resampled to match the resolution of the 96-direction acquisitions prior to analysis.

Cohort III consists of 8 healthy adult subjects (4 male, 4 female, ages 21 to 31 years) scanned for one to six sessions on scanner B (Figure 1). Each subject underwent all sessions within one year. Each subject received one to two scans during each session, with each scan consisting of a 40-direction b = 1000 s/mm^2 and a 56-direction b = 2000 s/mm^2 acquisition (Table 2). The majority of these scans were acquired at 2.1mm by 2.1mm by 2.2mm (sagittal, coronal, and axial) resolution and TE / TR = 79ms / 2900ms, with a few acquired at 2.5mm isotropic resolution and TE / TR = 75ms / 3000ms.

Cohort IV consists of 83 healthy child subjects (48 male, 35 female, ages 5 to 8 years) scanned for one to two sessions on scanner B (Figure 1). For the subjects with multiple sessions, the sessions were longitudinally acquired, spaced approximately one year apart. As with Cohort III, during each session, each subject received one to two scans, with each scan consisting of a 40-direction b = 1000 s/mm^2 and a 56-direction b = 2000 s/mm^2 acquisition (Table 2). These scans were acquired at 2.1mm by 2.1mm by 2.2mm (sagittal, coronal, and axial) resolution with TE / TR = 79ms / 2900ms.

All acquisitions were phase encoded in the posterior to anterior direction (APP) and were acquired with one b = 0 s/mm^2 volume each. Reverse phase encoded (APA) b = 0 s/mm^2 volumes were also acquired for all scans in all cohorts except for those from one subject in cohort II at site 3. Most sessions also included a T1-weighted image for structural analysis or distortion correction (36). All images were deidentified and all scans were acquired only after informed consent under supervision of the project Institutional Review Board.

Data preprocessing

After acquisition, all scans in MASiVar were preprocessed and quality checked with the PreQual pipeline (37). In brief, all acquisitions per scan were denoised with the Marchenko-Pastur technique (38–40), intensity normalized, and distortion corrected. Distortion correction included susceptibility-induced distortion correction (41) using APA b = 0 s/mm^2 volumes when available and the Synb0-DisCo deep learning framework (36) and associated T1 image

when not, eddy current-induced distortion correction, intervolume motion correction, and slice-wise signal drop out imputation (42,43). The estimated volume-to-volume displacement corrected during preprocessing and signal-to-noise ratios of the scans are reported in Supporting Information Figure S1.

Overview of variability study

Using data acquired in adults, we sought to demonstrate the capacity of MASiVar to simultaneously investigate DWI variability due to

1. intrasession (scans acquired within the same session on the same scanner of the same subject),
2. intersession (scans acquired between different sessions on the same scanner of the same subject),
3. interscanner (scans acquired between different sessions on different scanners of the same subject), and
4. intersubject (scans acquired of different subjects in different sessions on the same scanner) effects.

We quantified these levels of effects in four common types of DWI analysis, including

1. a diffusion tensor imaging (DTI) signal representation,multi-compartment neurite orientation dispersion and density imaging (NODDI) model (4),
2. the RecoBundles white matter bundle segmentation technique (44), and
3. a connectomics representation with graph-based measures (45).

For DTI, we investigate variability in regional fractional anisotropy (FA), mean diffusivity (MD), and principal eigenvector (V1) measurements. For NODDI, we investigate variability in regional cerebrospinal fluid (CSF) volume fraction (cVF), intracellular volume fraction (iVF), and orientation dispersion index (ODI) measurements. For bundle segmentation, we investigate variability in bundle shape, volume, length, and FA. For connectomics we investigate whole connectome variability as well as that of the maximum modularity (MM), global efficiency (GE), and characteristic path length (CPL) graph measures (Figure 2).

Defining variability

Due to the nested nature of session, scanner, and subject effects, we wished to define variability for a given effect in a way that limited confounding from the other effects. For example, because sessions are necessarily nested in scanners, we wanted to be able to investigate interscanner effects without confounding from intrasession effects and vice versa. To do this, we elected to use a paired difference approach. By identifying pairs of images that should produce the same measurements, we can estimate variability by analyzing the differences between the scans in the pair. For example, when the same person is scanned once on two different scanners, interscanner variability can be quantified by computing the differences in measurements between the two images. Additionally, since a pair of scans can only satisfy the criteria for exactly one of the listed effects, computing variability within these pairs

reduces confounding by holding the other effects constant. As such, we identified all pairs of scans in cohorts I to III of MASiVar that satisfied the criteria for each of the four listed effects (Figure 3), resulting in 41 intrasession pairs, 188 intersession pairs, 53 interscanner pairs, and 80 intersubject pairs. Only cohort II was used for the intersubject pairings to reduce bias toward scanners A and B.

We quantify variability at a given level of variation and type of DWI measurement by "summarizing" the "differences" within the relevant pairs. For example, to compute the variability of intersession DTI FA measurements, we compute the "difference" in FA in each intersession scan pair and report a "summary" across all pairs. Due to the distinct properties of each type of DWI processing, the exact definition of "difference" (Figure 4) and "summary" (Figure 5) vary by measurement type and are detailed in the following sections.

Variability in DTI and NODDI

For the DTI approach, we extract the b = 1000 s/mm^2 acquisition from each scan with the largest number of directions. We then calculate the diffusion tensor for each scan using an iteratively reweighted least squares approach implemented in MRtrix3 (46). The tensors are subsequently converted to FA, MD, and V1 representations of the data (47). These images are then deformably registered to the Montreal Neurological Institute (MNI) image space with the ANTs software package (48,49). From there, we identify the 48 regions of interest (ROIs) in each image defined by the Johns Hopkins white matter atlas (50–52) (Figure 2a).

For the NODDI approach, we extract the b = 1000 s/mm^2 acquisition from each scan with the largest number of directions and the b = 2000 s/mm^2 acquisition. We then fit the multicompartment model with the UCL NODDI Toolbox as implemented in MATLAB (4). The models are subsequently converted to cVF, iVF, and ODI representations. These images are then deformably registered to MNI space with the ANTs software package. From there, we identify the 48 ROIs in each image defined by the Johns Hopkins white matter atlas (Figure 2b).

We perform the DTI and NODDI paired difference calculations on a regional basis in MNI space with voxel-wise correspondence between images. For a given region and level of variation, we calculate the difference within a pair of scans for FA, MD, cVF, iVF, and ODI as the median voxel-wise absolute percent difference. For V1, we define it as the median voxel-wise absolute angular difference in degrees (Figure 4a). Once the paired differences are computed for all pairs in all regions, we compute the regional medians and take the resulting distribution across regions to summarize a given level of variability (Figure 5a). Additionally, we compute the 95% confidence intervals for each of the regional medians with nonparametric bootstrapping of the scan pairs with 1000 iterations (Supporting Information Figures S2-S7) (53).

Variability in bundle segmentation

For the white matter segmentation approach, we extract the b = 2000 s/mm^2 acquisition from each scan. We calculate a whole-brain tractogram with DIPY of 2 million streamlines (54). We use the constrained spherical deconvolution model (55) with probabilistic local tracking with a maximum angle of 25°, a seeding criterion of FA > 0.3, and a stopping criterion of FA < 0.2. We extract 43 white matter bundles (Supporting Information Table S1) from each tractogram using the RecoBundles algorithm as implemented in DIPY. In short, each tractogram is registered to an MNI tractogram template and streamlines from each tractogram are assigned to bundles within the template (44). The length, volume, and FA of each bundle are then calculated. We calculate bundle length by calculating the median streamline length. We calculate volume by first converting each bundle to a tract density image representation. From there, a binary bundle mask is calculated by thresholding the tract density image at 5% of the 99th percentile density. Volume is calculated by multiplying the number of voxels in the mask by the volume of each voxel. FA is calculated by first converting the image to a tensor representation (46) and then to an FA representation (47). Each bundle's binary mask is then applied to obtain the voxel-wise median FA value per bundle (Figure 2c).

Because streamline-wise and subsequent voxel-wise correspondence cannot be achieved with tractography and bundle segmentation, we compute paired differences differently than in the DTI and NODDI case. For a given bundle and level of variation, we calculate the paired difference of bundle shape with the Dice similarity index (56) between the tract density images from the two images. We define the difference of bundle volumes as the absolute percent difference. We calculate the difference of bundle FA as the absolute percent difference between the voxel-wise medians from each image and the difference of bundle length as the absolute percent difference between the streamline-wise medians from each image (Figure 4b). Similar to the DTI and NODDI analysis, once the paired differences are computed for all pairs in all bundles, we compute the bundle-wise medians and take the resulting distribution across bundles to summarize a given level of variability (Figure 5a). Additionally, we compute the 95% confidence intervals for each of the bundle-wise medians with nonparametric bootstrapping of the scan pairs with 1000 iterations (Supporting Information Figures S8-S11).

Variability in connectomics

For the connectomics approach, we extract the b = 2000 s/mm^2 acquisition from each scan. We then calculate a whole-brain tractogram with MRtrix3 (57). We first use the constrained spherical deconvolution model with probabilistic tracking with a maximum angle of 25°, a seeding criterion of FA > 0.3 and a stopping criterion of FA

< 0.2 to calculate a 10 million streamline tractogram. The tractogram is then filtered with the SIFT approach to 2 million streamlines (58). We parcellate the brain into 96 cortical regions using the Harvard-Oxford cortical atlas (59–62) and compute a connectome where each edge represents the average streamline distance connecting the two nodes. The MM, GE, and CPL are then calculated from each connectome using the Brain Connectivity Toolbox as implemented in MATLAB (45) (Figure 2d).

To evaluate paired differences of connectomics, we characterize each pair of connectomes as both (1) a whole and (2) through scalar measures. First, we calculate the Pearson correlation between the connectomes within each image pair as an estimate for connectome agreement within a pair. Second, we calculate the percent absolute difference in MM, GE, and CPL (45) between the connectomes in the pair (Figure 4c). Unlike the DTI, NODDI, and bundle segmentation cases, we do not have multiple regions or bundles for the connectomics analysis with which we can obtain a distribution to summarize variability at a level. As a result, we instead compute a distribution for the median paired difference of each metric with nonparametric bootstrapping of the scan pairs with 1000 iterations (Figure 5b).

Comparing variability across levels

To compare variability across the intrasession, intersession, interscanner, and intersubject levels, we use non-parametric statistical tests for all four processing approaches with a significance level of 0.05 and report the uncorrected *p*-values. For the DTI, NODDI, and bundle segmentation analyses, we use the Wilcoxon signed-rank test for paired distributions, as all points in each distribution are measured from corresponding regions or bundles (63). For the connectomics analysis, we use the Wilcoxon rank-sum test for unpaired distributions, as the distributions were constructed from bootstrapping and thus do not have correspondence (63).

Comparing variability across processing approaches

Last, to obtain a more global understanding of the relative intrasubject (intrasession, intersession, and interscanner) effects across different processing approaches, we normalize these effects to intersubject variation for each approach. We calculate the ratios of each of the intrasubject differences to the corresponding intersubject estimates. For the DTI, NODDI, and bundle segmentation analysis, we compute this ratio on a region- or bundle-wise basis, and for the connectome analysis, we compute this ratio on the bootstrapped distributions. For the Dice and correlation similarity measures, we first subtract them from 1 to obtain dissimilarity estimates. With this design, ratios of <1, 1, and >1 indicate the intrasubject variation is less than, equal to, or greater than the intersubject variation, respectively.

RESULTS

Variability in DTI

As shown in Figure 6, we find that intrasession FA measurements vary around 8.3%, that intersession measurements vary around 9.1%, that interscanner measurements vary around 13.7%, and that intersubject measurements vary around 17.7%. We find the corresponding measurements in the MD case to be 3.6%, 4.9%, 8.6%, and 9.1% and for the V1 case to be 10.0°, 10.7°, 14.5°, and 20.2°, respectively. All these differences were statistically significant ($p < 0.0005$, Wilcoxon signed-rank test), except for the interscanner and intersubject comparison for MD ($p < 0.05$).

Variability in NODDI

As shown in Figure 7, we find that the intrasession cVF measurements vary around 47.3%, that intersession measurements vary around 55.1%, that interscanner measurements vary around 57.5%, and that intersubject measurements vary around 61.9%. We find the corresponding measurements in the iVF case to be 7.9%, 8.6%, 11.8%, and 12.1% and for the ODI case to be 11.7%, 13.5%, 19.1%, and 26.5%, respectively. All these differences were statistically significant ($p < 0.0005$, Wilcoxon signed-rank test), except for the interscanner and intersubject iVF comparison. We evaluated cVF only in white matter regions defined by the Johns Hopkins atlas and thus dealt with very low cVF values when calculating percent difference.

Variability in bundle segmentation

As shown in Figure 8, we find that intrasession bundles overlap at 0.61 Dice, that intersession bundles overlap at 0.61 Dice, that interscanner bundles overlap at 0.57 Dice, and that intersubject bundles overlap at 0.53 Dice. We find the measurements for the corresponding levels of variation in the bundle volume comparisons to be 11.4%, 10.2%, 12.8%, and 17.9%, in the FA case to be 1.7%, 1.9%, 5.0%, and 4.9%, and in the bundle length case to be 3.2%, 3.6%, 4.1%, and 8.0%, respectively. All these differences were statistically significant ($p < 0.0005$, Wilcoxon signed-rank test), except for the intrasession and intersession shape, volume, and length ($p < 0.05$) comparisons as well as the interscanner and intersubject FA comparison.

Variability in connectomics

As shown in Figure 9, we find that the intrasession connectomes correlate at 0.62, that the intersession connectomes correlate at 0.59, that the interscanner connectomes correlate at 0.59 and that the intersubject connectomes correlate at 0.54. We find the measurements for the corresponding levels of variation in the MM case to be 7.2%, 8.2%, 9.5%, and 13.3%, in the GE case to be 1.0%, 1.6%, 6.1%, and 6.2%, and in the CPL case to be 0.8%, 1.5%, 6.2%, and 6.1%, respectively. All these differences were statistically significant ($p < 0.0005$, Wilcoxon rank-sum test) except for the interscanner and intersubject comparison of CPL.

Comparing variability across processing approaches

We plot the ratios of intrasubject to intersubject variability in Figure 10. Previous studies have demonstrated what we find to be good intrasession and intersession FA reproducibility with coefficients of variation (CoV) up to 3% as well as moderate interscanner reproducibility with CoV up to 8% (Table 1). Based on these findings, we draw thresholds to separate the FA intrasession and intersession ratios from the interscanner ratios and the FA interscanner ratios from the intersubject reference in order to compare the remaining metrics. As such, we take a ratio of <0.65 to represent good reproducibility, 0.65-0.95 to represent moderate, and >0.95 to represent poor reproducibility. With this heuristic, we find good to moderate reproducibility in most of the intrasession and intersession measures across all four approaches. One notable deviation from this trend is that intersession white matter cVF demonstrated moderate to poor reproducibility. Additionally, despite the thresholds drawn, we note that the intrasession and intersession variability estimates are all within an order of magnitude of the intersubject effects as designated by ratios between 0.1 and 1. Last, for most measures, we find moderate to poor interscanner reproducibility and that interscanner effects can introduce variability that is on par with that from intersubject effects. We find this trend is most striking for the GE and CPL analyses that demonstrated intrasession ratios of around 0.2 but interscanner ratios near 1. In summary, we find that consistently across all four DWI analysis approaches intrasubject variability is good to moderate but non-negligible compared to intersubject variability and that the size of interscanner effects can approach the size of intersubject effects.

DISCUSSION AND CONCLUSIONS

Here, we present, MASiVar, a dataset designed for investigation of DWI variability. Additionally, to demonstrate the capacity of MASiVar as a resource, we characterize intrasession, intersession, interscanner, and intersubject variability and plot the ratios of intrasubject to intersubject variability in four common diffusion processing approaches. In support of our hypothesis, we consistently find that variability increases with consideration of session, scanner, and subject effects. We also consistently find across all approaches that intrasubject variability accounts for a non-negligible portion of intersubject variability and that at times interscanner variability can

approach intersubject variability. We interpret two primary conclusions from these results. The first is that MASiVar provides the field a resource to obtain an improved global understanding of session, scanner, and subject effects within and between different DWI processing approaches. Second, we interpret these results to mean that harmonization between scanners for multisite analyses should be carefully considered prior to inference of group differences on subjects.

The reproducibility of DWI analyses has received significant attention in the field, including the analysis of tensor representations (23–26), multi-compartment models (26,27), tractography and bundle segmentation (28,29), and connectomics (33,35) (Table 1). Looking at the literature, we find similar trends between our results and those of prior studies that variation increases with session and scanner biases, although the numbers are not directly comparable due to different definitions of variability. However, review of the literature also demonstrates a fragmented picture of DWI variability. Previous studies have largely each primarily focused on one type of approach and one or two levels of variation. This coupled with the different definitions of variability and different study objectives have made it difficult to understand how the different effects relate to each other and how they affect a multitude of common DWI processing approaches. To the best of our knowledge, this study represents the first attempt and to characterize all four types of diffusion processing and all four levels of variation consistently and simultaneously. Thus, we hope that the dataset and study presented here will promote further investigation into a wide spectrum of DWI variability issues from a large pool of models to push the field toward a global understanding of the effects of session, scanner, and subject biases on different DWI measurements.

Of note, the bulk of existing studies used CoV to estimate variation and the intraclass correlation coefficient (ICC) to characterize the proportion of intersubject variation attributable to session or scanner effects. In contrast, we used an intrasubject and intersubject paired difference approach for the former and quantified their ratios for the latter. We chose these approaches to improve isolation of session and scanner effects. For instance, by pairing images, we remove other potential confounders like subject or acquisition effects. This approach also has the benefit of characterizing the relative sizes of intra- and intersubject variation without the necessary assumptions of traditional ICC, including not having additional confounders within each class, an assumption that is inherently violated when looking at sessions nested within scanners nested with subjects.

For this study, we chose popular software toolboxes to do all the analyses, parameter configurations that we were familiar with, and consistent similarity assessments that we found to be interpretable. However, we recognize that there are many other software options available to do similar tasks, each with a large number of different configurations, and a large number of ways to assess variability. For instance, there are different methods for fitting tensors (64–66), for identifying regions (61,67–69) and bundles (70–73), for comparing bundles (74), and for

configuring and representing connectomes (35,45,75,76). Additionally, there are a number of other microstructural measures that can be characterized as well (19). Thus, the goal of the present study was not to provide an analysis between different processing toolboxes or parameters, and since each approach was not necessarily optimized, we do not recommend thorough utilization of absolute reproducibility values presented here for any one processing approach. Instead, we aimed to contribute to a global understanding of DWI variability and its relative trends across the four processing approaches and across sessions, scanners, and subjects in a generally interpretable way that demonstrated the potential of the dataset. As such, we hope that the release of MASiVar will prompt other investigators in the field to optimize and further characterize differences between software tools and their parameters, different DWI processing and variability measures, and other potential confounders in DWI analysis.

In addition to the ability of MASiVar to serve as a utility for variability analysis, we note that the pediatric subjects in cohort IV present another unique resource for the field. The majority of the existing DWI datasets and studies for variability use adult subjects. Of existing pediatric datasets, many have focused on older age ranges. For example, the Adolescent Brain Cognitive Development project (77) and the Lifespan Human Connectome Project in Development (78) contain longitudinal DWI data acquired from children starting at age 9 and 10 through adolescence. Thus, to the best of our knowledge, MASiVar represents one of the first publicly available longitudinal DWI datasets of children prior to adolescence aged 5-8 years old and is further distinguished by its inclusion of repeated scans within each session. As a demonstration of the usefulness of cohort IV, we include a characterization of the longitudinal intersession variability in children with one year between sessions compared to the adult intersession variability computed above for all four processing approaches (Supporting Information Figure S12). We find that, as expected, the pediatric cohort exhibits increased variability for many measurements, supporting that the expected developmental changes in this age group likely introduce additional variability on top of session effects. We hope that investigators in developmental neuroscience and pediatric neurology will be able to take advantage of this resource for their work.

One limitation of the variability study is the differences in number of gradient directions between the different cohorts. Cohort III consists of a 40-direction b = 1000 s/mm^2 acquisition and a 56-direction b = 2000 s/mm^2 acquisition in contrast to the 96 directions for cohorts I and II. There is a potential effect that could be biasing the results. However, our study design focuses on paired scans and thus all variability measurements are between acquisitions of the same configuration, which should minimize this effect. In a similar vein, due to hardware limitations, the data collected at site 3 in cohort II was collected at a maximum shell of 2465 s/mm^2 as opposed to the 2500 s/mm^2 across the rest of MASiVar. This shell was not used for the present variability analysis, but this discrepancy should be noted on future studies using the dataset.

Last, we have made the MASiVar dataset publicly available at https://openneuro.org/datasets/ds003416 in Brain Imaging Data Structure (BIDS) format (79) with deidentified metadata and defaced images.

ACKNOWLEDGEMENTS

The authors thank E. Brian Welch for his help with image acquisition and study design, Zachary J. Williams for his insight into bootstrapping, and the reviewers for their thoughtful and critical feedback in improving this manuscript. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. This work was supported by the National Institutes of Health (NIH) under award numbers 5R01EB017230, 5T32EB001628, 5T32GM007347, and 1UL1RR024975. This work was also supported by the National Science Foundation (NSF) under award numbers 1452485, 1660816, and 1750213. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

REFERENCES

1. O'Donnell LJ, Westin CF. An introduction to diffusion tensor image analysis. Neurosurg. Clin. N. Am. 2011;22:185–196 doi: 10.1016/j.nec.2010.12.004.
2. Assaf Y, Pasternak O. Diffusion tensor imaging (DTI)-based white matter mapping in brain research: A review. J. Mol. Neurosci. 2008;34:51–61 doi: 10.1007/s12031-007-0029-0.
3. Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophys. J. 1994;66:259–267 doi: 10.1016/S0006-3495(94)80775-1.
4. Zhang H, Schneider T, Wheeler-Kingshott CA, Alexander DC. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 2012;61:1000–1016 doi: 10.1016/j.neuroimage.2012.03.072.
5. Stanisz GJ, Szafer A, Wright GA, Henkelman RM. An analytical model of restricted diffusion in bovine optic nerve. Magn. Reson. Med. 1997;37:103–111 doi: 10.1002/mrm.1910370115.
6. Panagiotaki E, Schneider T, Siow B, Hall MG, Lythgoe MF, Alexander DC. Compartment models of the diffusion MR signal in brain white matter: A taxonomy and comparison. Neuroimage 2012;59:2241–2254 doi: 10.1016/j.neuroimage.2011.09.081.
7. Schilling KG, Petit L, Rheault F, et al. Brain connections derived from diffusion MRI tractography can be highly anatomically accurate—if we know where white matter pathways start, where they end, and where they do not go. Brain Struct. Funct. 2020;225:2387–2402 doi: 10.1007/s00429-020-02129-z.
8. Sotiropoulos SN, Zalesky A. Building connectomes using diffusion MRI: why, how and but. NMR Biomed.

2019;32:1–23 doi: 10.1002/nbm.3752.
9. Di Martino A, O'Connor D, Chen B, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci. Data 2017;4:170010 doi: 10.1038/sdata.2017.10.
10. Travers BG, Adluru N, Ennis C, et al. Diffusion Tensor Imaging in Autism Spectrum Disorder: A Review. Autism Res. 2012;5:289–313 doi: 10.1002/aur.1243.
11. Repple J, Karliczek G, Meinert S, et al. Variation of HbA1c affects cognition and white matter microstructure in healthy, young adults. Mol. Psychiatry 2019:1–10 doi: 10.1038/s41380-019-0504-3.
12. Kodl CT, Franc DT, Rao JP, et al. Diffusion tensor imaging identifies deficits in white matter microstructure in subjects with type 1 diabetes that correlate with reduced neurocognitive function. Diabetes 2008;57:3083–3089 doi: 10.2337/db08-0724.
13. De Santis S, Bastiani M, Droby A, et al. Characterizing Microstructural Tissue Properties in Multiple Sclerosis with Diffusion MRI at 7 T and 3 T: The Impact of the Experimental Design. Neuroscience 2019;403:17–26 doi: 10.1016/j.neuroscience.2018.03.048.
14. Cetin-Karayumak S, Di Biase MA, Chunga N, et al. White matter abnormalities across the lifespan of schizophrenia: a harmonized multi-site diffusion MRI study. Mol. Psychiatry 2019 doi: 10.1038/s41380-019-0509-y.
15. Westlye LT, Walhovd KB, Dale AM, et al. Life-span changes of the human brain white matter: Diffusion tensor imaging (DTI) and volumetry. Cereb. Cortex 2010;20:2055–2068 doi: 10.1093/cercor/bhp280.
16. Menzler K, Belke M, Wehrmann E, et al. Men and women are different: Diffusion tensor imaging reveals sexual dimorphism in the microstructure of the thalamus, corpus callosum and cingulum. Neuroimage 2011;54:2557–2562 doi: 10.1016/j.neuroimage.2010.11.029.
17. Froeling M, Tax CMW, Vos SB, Luijten PR, Leemans A. "MASSIVE" brain dataset: Multiple acquisitions for standardization of structural imaging validation and evaluation. Magn. Reson. Med. 2017;77:1797–1809 doi: 10.1002/mrm.26259.
18. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: An overview. Neuroimage 2013;80:62–79 doi: 10.1016/j.neuroimage.2013.05.041.
19. Koller K, Rudrapatna SU, Chamberland M, et al. MICRA: Microstructural Image Compilation with Repeated Acquisitions. Neuroimage 2020;225:117406 doi: 10.1016/j.neuroimage.2020.117406.
20. Duchesne S, Chouinard I, Potvin O, et al. The Canadian Dementia Imaging Protocol: Harmonizing National Cohorts. J. Magn. Reson. Imaging 2019;49:456–465 doi: 10.1002/jmri.26197.
21. Tong Q, He H, Gong T, et al. Multicenter dataset of multi-shell diffusion MRI in healthy traveling adults with identical settings. Sci. Data 2020;7 doi: 10.1038/s41597-020-0493-8.
22. Farrell JAD, Landman BA, Jones CK, et al. Effects of SNR on the Accuracy and Reproducibility of DTI-derived Fractional Anisotropy, Mean Diffusivity, and Principal Eigenvector Measurements at 1.5T. J. Magn.

Reson. 2010;26:756–767 doi: 10.1002/jmri.21053.Effects.
23. Magnotta VA, Matsui JT, Liu D, et al. Multicenter Reliability of Diffusion Tensor Imaging. Brain Connect. 2012;2:345–355 doi: 10.1089/brain.2012.0112.
24. Vollmar C, O'Muircheartaigh J, Barker GJ, et al. Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. Neuroimage 2010;51:1384–1394 doi: 10.1016/j.neuroimage.2010.03.046.
25. Palacios EM, Martin AJ, Boss MA, et al. Towards Precision and Reproducibility of DTI. AJNR Am J Neuroradiol. 2017;38:537–545 doi: 10.3174/ajnr.A5025.Towards.
26. Andica C, Kamagata K, Hayashi T, et al. Scan–rescan and inter-vendor reproducibility of neurite orientation dispersion and density imaging metrics. Neuroradiology 2020;62:483–494 doi: 10.1007/s00234-019-02350-6.
27. Tariq M, Schneider T, Alexander DC, Wheeler-Kingshott CA, Zhang H. Assessing scan-rescan reproducibility of the parameter estimates from NODDI. Proc. Intl. Soc. Mag. Reson. Med. 21 2013;21:3187.
28. Nath V, Schilling KG, Parvathaneni P, et al. Tractography reproducibility challenge with empirical data (TraCED): The 2017 ISMRM diffusion study group challenge. J. Magn. Reson. Imaging 2020;51:234–249 doi: 10.1002/jmri.26794.
29. Besseling RMH, Jansen JFA, Overvliet GM, et al. Tract specific reproducibility of tractography based morphology and diffusion metrics. PLoS One 2012;7 doi: 10.1371/journal.pone.0034125.
30. Cousineau M, Jodoin PM, Morency FC, et al. A test-retest study on Parkinson's PPMI dataset yields statistically significant white matter fascicles. NeuroImage Clin. 2017;16:222–233 doi: 10.1016/j.nicl.2017.07.020.
31. Boukadi M, Marcotte K, Bedetti C, et al. Test-retest reliability of diffusion measures extracted along white matter language fiber bundles using Hardi-based tractography. Front. Neurosci. 2019;13 doi: 10.3389/fnins.2018.01055.
32. Schilling KG, Rheault F, Petit L, Hansen CB, Nath V, Yeh F. Tractography dissection variability : what happens when 42 groups dissect 14 white matter bundles on the same dataset ? bioRxiv 2020.
33. Prčkovska V, Rodrigues P, Puigdellivol Sanchez A, et al. Reproducibility of the Structural Connectome Reconstruction across Diffusion Methods. J. Neuroimaging 2016;26:46–57 doi: 10.1111/jon.12298.
34. Girard G, Whittingstall K, Deriche R, Descoteaux M. Studying white matter tractography reproducibility through connectivity matrices. Int. Symp. Magn. Reson. Med. 2015.
35. Roine T, Jeurissen B, Perrone D, et al. Reproducibility and intercorrelation of graph theoretical measures in structural brain connectivity networks. Med. Image Anal. 2019;52:56–67 doi: 10.1016/j.media.2018.10.009.
36. Schilling KG, Blaber J, Hansen C, et al. Distortion correction of diffusion weighted MRI without reverse phase-encoding scans or field-maps. PLoS One 2020;15:e0236418.
37. Cai LY, Yang Q, Hansen CB, et al. PreQual: An automated pipeline for integrated preprocessing and quality

assurance of diffusion weighted MRI images. Magn. Reson. Med. 2021:1–15 doi: 10.1002/mrm.28678.
38. Veraart J, Novikov DS, Christiaens D, Ades-aron B, Sijbers J, Fieremans E. Denoising of diffusion MRI using random matrix theory. Neuroimage 2016;142:394–406 doi: 10.1016/j.neuroimage.2016.08.016.
39. Veraart J, Fieremans E, Novikov DS. Diffusion MRI noise mapping using random matrix theory. Magn. Reson. Med. 2016;76:1582–1593 doi: 10.1002/mrm.26059.
40. Cordero-Grande L, Christiaens D, Hutter J, Price AN, Hajnal J V. Complex diffusion-weighted image estimation via matrix recovery under general noise models. Neuroimage 2019;200:391–404 doi: 10.1016/j.neuroimage.2019.06.039.
41. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. Neuroimage 2003;20:870–888 doi: 10.1016/S1053-8119(03)00336-7.
42. Andersson JLR, Sotiropoulos SN. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage 2016;125:1063–1078 doi: 10.1016/j.neuroimage.2015.10.019.
43. Andersson JLR, Graham MS, Zsoldos E, Sotiropoulos SN. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. Neuroimage 2016;141:556–572 doi: 10.1016/j.neuroimage.2016.06.058.
44. Garyfallidis E, Côté MA, Rheault F, et al. Recognition of white matter bundles using local and global streamline-based registration and clustering. Neuroimage 2018;170:283–295 doi: 10.1016/j.neuroimage.2017.07.015.
45. Rubinov M, Sporns O. Complex network measures of brain connectivity: Uses and interpretations. Neuroimage 2010;52:1059–1069 doi: 10.1016/j.neuroimage.2009.10.003.
46. Veraart J, Sijbers J, Sunaert S, Leemans A, Jeurissen B. Weighted linear least squares estimation of diffusion MRI parameters: Strengths, limitations, and pitfalls. Neuroimage 2013;81:335–346 doi: 10.1016/j.neuroimage.2013.05.028.
47. Westin CF, Peled S, Gudbjartsson H, Kikinis R, Jolesz FA. Geometrical Diffusion Measures for MRI from Tensor Basis Analysis. Proc. 5th Annu. Meet. ISMRM 1997:1742.
48. Tustison NJ, Cook PA, Klein A, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage 2014;99:166–179 doi: 10.1016/j.neuroimage.2014.05.044.
49. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 2008;12:26–41 doi: 10.1016/j.media.2007.06.004.
50. Mori S, Wakana S, Van Zijl PCM, Nagae-Poetscher LM. MRI atlas of human white matter. Elsevier; 2005.
51. Wakana S, Caprihan A, Panzenboeck MM, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. Neuroimage 2007;36:630–644 doi: 10.1016/j.neuroimage.2007.02.049.

52. Hua K, Zhang J, Wakana S, et al. Tract probability maps in stereotaxic spaces: Analyses of white matter anatomy and tract-specific quantification. Neuroimage 2008;39:336–347 doi: 10.1016/j.neuroimage.2007.07.053.
53. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. 1986;1:54–75 doi: 10.1214/ss/1177013815.
54. Garyfallidis E, Brett M, Amirbekian B, et al. Dipy, a library for the analysis of diffusion MRI data. Front. Neuroinform. 2014;8:1–17 doi: 10.3389/fninf.2014.00008.
55. Tournier JD, Calamante F, Connelly A. Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. Neuroimage 2007;35:1459–1472 doi: 10.1016/j.neuroimage.2007.02.016.
56. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.
57. Tournier J-D, Smith R, Raffelt D, et al. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. 2019 doi: 10.1016/j.neuroimage.2019.116137.
58. Smith RE, Tournier JD, Calamante F, Connelly A. SIFT: Spherical-deconvolution informed filtering of tractograms. Neuroimage 2013;67:298–312 doi: 10.1016/j.neuroimage.2012.11.049.
59. Makris N, Goldstein JM, Kennedy D, et al. Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr. Res. 2006;83:155–171 doi: 10.1016/j.schres.2005.11.020.
60. Frazier JA, Chiu S, Breeze JL, et al. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. Am. J. Psychiatry 2005;162:1256–1265 doi: 10.1176/appi.ajp.162.7.1256.
61. Desikan RS, Ségonne F, Fischl B, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006;31:968–980 doi: 10.1016/j.neuroimage.2006.01.021.
62. Goldstein JM, Seidman LJ, Makris N, et al. Hypothalamic Abnormalities in Schizophrenia: Sex Effects and Genetic Vulnerability. Biol. Psychiatry 2007;61:935–945 doi: 10.1016/j.biopsych.2006.06.027.
63. Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods. John Wiley & Sons; 2013.
64. Hernandez-Fernandez M, Reguly I, Jbabdi S, Giles M, Smith S, Sotiropoulos SN. Using GPUs to accelerate computational diffusion MRI: From microstructure estimation to tractography and connectomes. Neuroimage 2019;188:598–615 doi: 10.1016/j.neuroimage.2018.12.015.
65. Chang LC, Jones DK, Pierpaoli C. RESTORE: Robust estimation of tensors by outlier rejection. Magn. Reson. Med. 2005;53:1088–1095 doi: 10.1002/mrm.20426.
66. Cook P a, Bai Y, Seunarine KK, Hall MG, Parker GJ, Alexander DC. Camino: Open-Source Diffusion-MRI Reconstruction and Processing. 14th Sci. Meet. Int. Soc. Magn. Reson. Med. 2006;14:2759.
67. Hansen CB, Yang Q, Lyu I, et al. Pandora: 4-D white matter bundle population-based atlases derived from diffusion MRI fiber tractography. bioRxiv 2020:2020.06.12.148999 doi: 10.1101/2020.06.12.148999.
68. Volz LJ, Cieslak M, Grafton ST. A probabilistic atlas of fiber crossings for variability reduction of anisotropy

measures. Brain Struct. Funct. 2018;223:635–651 doi: 10.1007/s00429-017-1508-x.
69. Figley TD, Mortazavi Moghadam B, Bhullar N, Kornelsen J, Courtney SM, Figley CR. Probabilistic white matter atlases of human auditory, basal ganglia, language, precuneus, sensorimotor, visual and visuospatial networks. Front. Hum. Neurosci. 2017;11:1–12 doi: 10.3389/fnhum.2017.00306.
70. Warrington S, Bryant KL, Khrapitchev AA, et al. XTRACT - Standardised protocols for automated tractography in the human and macaque brain. Neuroimage 2020;217:1–15 doi: 10.1016/j.neuroimage.2020.116923.
71. Yeh FC. Shape analysis of the human association pathways. Neuroimage 2020;223 doi: 10.1016/j.neuroimage.2020.117329.
72. Yeh FC, Verstynen TD, Wang Y, Fernández-Miranda JC, Tseng WYI. Deterministic diffusion fiber tracking improved by quantitative anisotropy. PLoS One 2013;8:1–16 doi: 10.1371/journal.pone.0080713.
73. Yendiki A, Panneck P, Srinivasan P, et al. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. Front. Neuroinform. 2011;5:1–12 doi: 10.3389/fninf.2011.00023.
74. Rheault F, De Benedictis A, Daducci A, et al. Tractostorm: The what, why, and how of tractography dissection reproducibility. Hum. Brain Mapp. 2020;41:1859–1874 doi: 10.1002/hbm.24917.
75. Hagmann P, Cammoun L, Gigandet X, et al. Mapping the structural core of human cerebral cortex. PLoS Biol. 2008;6:1479–1493 doi: 10.1371/journal.pbio.0060159.
76. Sporns O, Tononi G, Kötter R. The human connectome: A structural description of the human brain. PLoS Comput. Biol. 2005;1:0245–0251 doi: 10.1371/journal.pcbi.0010042.
77. Casey BJ, Cannonier T, Conley MI, et al. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. Dev. Cogn. Neurosci. 2018;32:43–54 doi: 10.1016/j.dcn.2018.03.001.
78. Somerville LH, Bookheimer SY, Buckner RL, et al. The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5–21 year olds. Neuroimage 2018;183:456–468 doi: 10.1016/j.neuroimage.2018.08.050.
79. Gorgolewski KJ, Auer T, Calhoun VD, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 2016;3:160044 doi: 10.1038/sdata.2016.44.

Table 1. A survey of existing DWI variability estimates (cited with first author and publication year) against those presented in the present work (blue).

	Intrasession	Intersession	Interscanner	Intersubject
DTI				
FA	8.32% PD 2% CoV (Farrell, 2010) (22)	9.07% PD 1% CoV (Magnotta, 2012) (23) 1-2% CoV (Vollmar, 2010) (24) 3% CoV (Palacios, 2017) (25) 0.5% CoV (Andica, 2020) (26) 0.90-0.99 ICC (Vollmar, 2010) (24) 0.74-1.00 ICC (Andica, 2020) (26) 3% CoV (Farrell, 2010) (22) 0.6-1% CoV (Koller, 2020) (19) 0.93-0.97 ICC (Koller, 2020) (19)	13.73% PD 3% CoV (Magnotta, 2012) (23) 2-4% CoV (Vollmar, 2010) (24) 8% CoV (Palacios, 2017) (25) 2% CoV (Andica, 2020) (26) 0.82-0.99 ICC (Vollmar, 2010) (24) 0.54-0.97 ICC (Andica, 2020) (26)	17.66% PD
MD	3.63% PD 1% CoV (Farrell, 2010) (22)	4.92% PD 1% CoV (Magnotta, 2012) (23) 2% CoV (Palacios, 2017) (25) 0.2% CoV (Andica, 2020) (26) 1% CoV (Farrell, 2010) (22) 0.5-1% CoV (Koller, 2020) (19) 0.94-0.96 ICC (Koller, 2020) (19)	8.61% PD 2% CoV (Magnotta, 2012) (23) 6% CoV (Palacios, 2017) (25) 3% CoV (Andica, 2020) (26)	9.06% PD
V1	9.97°	10.72° 7-10° (Farrell, 2010) (22)	14.46°	20.19°
NODDI				
cVF	47.32% PD	55.07% PD 1.6-3.6% CoV (Andica, 2020) (26) 0.133-0.997 ICC (Andica, 2020) (26)	57.47% PD 15.9% CoV (Andica, 2020) (26) 0.013-0.545 ICC (Andica, 2020) (26)	61.93% PD
iVF	7.92% PD	8.61% PD 0.4% CoV (Andica, 2020) (26) 0.773-0.989 ICC (Andica, 2020) (26) 5.1% CoV (Tariq, 2013) (27)	11.76% PD 0.9% CoV (Andica, 2020) (26) 0.300-0.935 ICC (Andica, 2020) (26)	12.13% PD
ODI	11.74% PD	13.51% PD 0.2-0.3% CoV (Andica, 2020) (26) 0.789-0.998 ICC (Andica, 2020) (26) 5.7% CoV (Tariq, 2013) (27)	19.05% PD 4.2% CoV (Andica, 2020) (26) 0.181-0.962 ICC (Andica, 2020) (26)	26.48% PD
Bundle Segmentation				
Shape	0.61 Dice ~0.67 Dice (Nath, 2020) (28)	0.61 Dice ~0.64 Dice (Nath, 2020) (28) 0.65-0.92 Dice (Besseling, 2012) (29) 0.72 wDice (Cousineau, 2017) (30) 0.71-0.87 wDice (Boukadi, 2019) (31)	0.57 Dice ~0.58 Dice (Nath, 2020) (28)	0.53 Dice ~0.5-0.6 Dice (Schilling, 2020) (32)
Volume	11.38% PD	10.20% PD 3-22% CoV (Besseling, 2012) (29) 0.53-0.96 ICC (Besseling, 2012) (29) 0.41-0.83 ICC (Boukadi, 2019) (31)	12.76% PD	17.90% PD
FA	1.66% PD	1.91% PD 1-4% CoV (Besseling, 2012) (29) 0.65-0.94 ICC (Besseling, 2012) (29) 0.62-0.89 ICC (Boukadi, 2019) (31)	5.00% PD	4.93% PD
Length	3.18% PD	3.57% PD 0.68-0.89 ICC (Boukadi, 2019) (31)	4.05% PD	8.04% PD
Connectomics				
Correlation	0.62 PC	0.59 PC 0.6-0.95 PC (Prčkovska, 2016) (33)	0.59 PC	0.54 PC 32.7-39.9% CD (Girard, 2015) (34)
MM	7.25% PD	8.20% PD	9.55% PD	13.26% PD
GE	1.03% PD	1.65% PD 31% CoV (Roine, 2019) (35) 0.78 ICC (Roine, 2019) (35)	6.06% PD	6.24% PD
CPL	0.80% PD	1.53% PD 2% CoV (Roine, 2019) (35) 0.77 ICC (Roine, 2019) (35)	6.19% PD	6.14% PD

PD: paired difference
PC: Pearson correlation
CoV: coefficient of variation
ICC: intraclass correlation coefficient
wDice: weighted Dice
CD: connectome distance, $\Sigma|C_1-C_2|/2$, where C_1 and C_2 are matrices

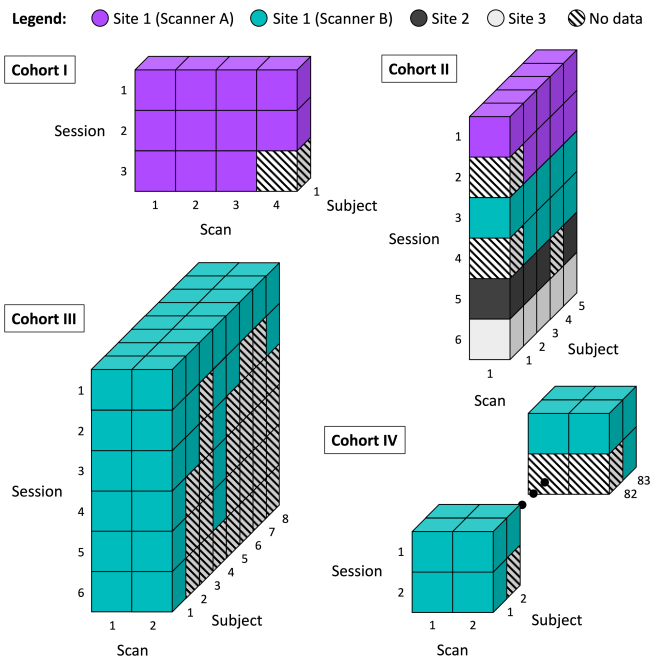

Figure 1. Overview of the MASiVar dataset. This dataset consists of four cohorts. Cohort I consists of one adult subject scanned repeatedly on one scanner. This subject underwent three separate imaging sessions and acquired 3-4 scans per session. Cohort II consists of 5 adult subjects each scanned on 3-4 different scanners across 3 institutions. Each subject underwent 1-2 sessions on each scanner and had one scan acquired per session. Cohort III consists of 8 adult subjects all scanned on one scanner. Each subject underwent 1-6 sessions on the scanner and had two scans acquired per session. Cohort IV consists of 83 child subjects all scanned on one scanner. Each subject underwent 1-2 sessions on the scanner and had two scans acquired per session.

Table 2. Acquisitions acquired in each scan for the different MASiVar cohorts.

Acquisitions Per Scan		
Cohort	*Shell (b-value)*	*Number of Directions*
I	1000	96
	1500	96
	2000	96
	2500	96
	3000	96
II	1000	30 or 32
	1000	96
	1500	96
	2000	96
	2465 or 2500	96
III	1000	40
	2000	56
IV	1000	40
	2000	56

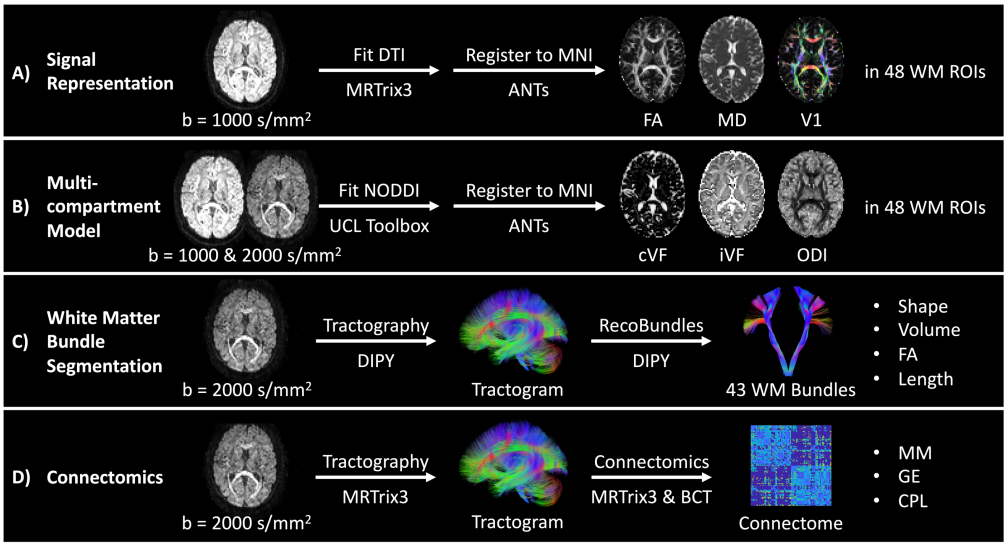

Figure 2. Outline of processing and measurements investigated presently in four common diffusion MRI analysis approaches. (A) We quantify variability in the tensor-based FA, MD, and V1 measurements in MNI space for each of the 48 Johns Hopkins white matter atlas regions. (B) We quantify variability in the NODDI-based cVF, iVF, and ODI measurements in MNI space for each of the 48 Johns Hopkins white matter atlas regions. (C) We quantify variability in bundle shape, volume, FA and length for 43 white matter bundles (Supporting Information Table S1) identified with the RecoBundles segmentation method. (D) We quantify variability in whole brain structural connectomes and the MM, GE, and CPL graph measures.

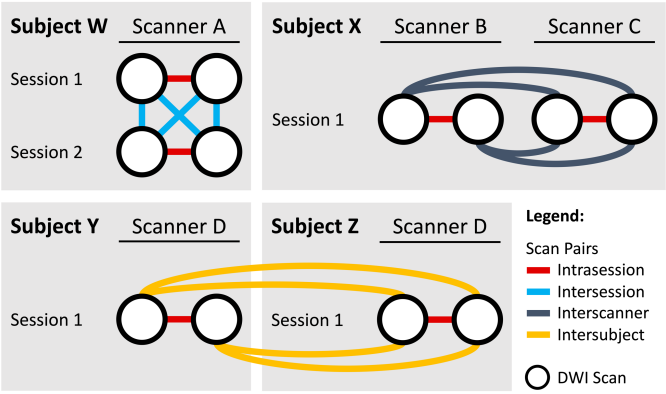

Figure 3. Example identification of scan pairs at the four levels of variation. The MASiVar dataset consists of scans that can be paired in order to satisfy intrasession, intersession, interscanner, and intersubject criteria. Each of these pairs represent scans that should produce the same measurements, thus quantification of differences within pairs provides an estimate of variability.

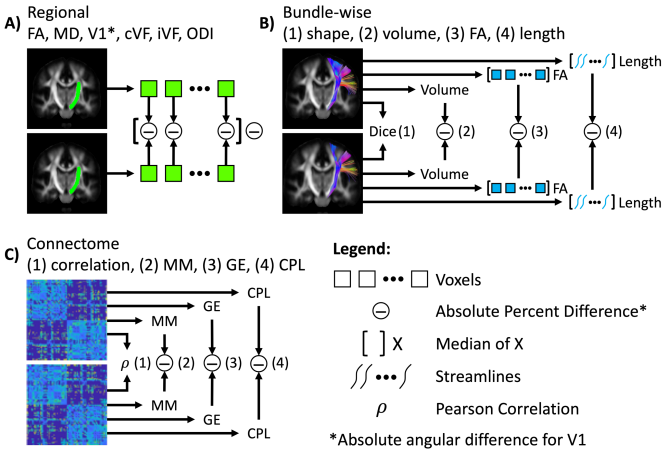

Figure 4. Defining variability within a pair of images. (A) Regional FA, MD, cVF, iVF, and ODI variability was defined as the median voxel-wise absolute percent difference. Absolute angular difference was used for V1. (B) Bundle-wise variability was defined as the Dice similarity for shape (1), as the absolute percent difference for volume (2), as the absolute percent difference in median voxel-wise FA for FA (3), and as the absolute percent difference in median streamline-wise length for length (4). (C) Whole connectome variability was defined as the Pearson correlation between connectivity matrices (1), and variability in the MM, GE, and CPL graph measures was defined with absolute percent difference (2-4).

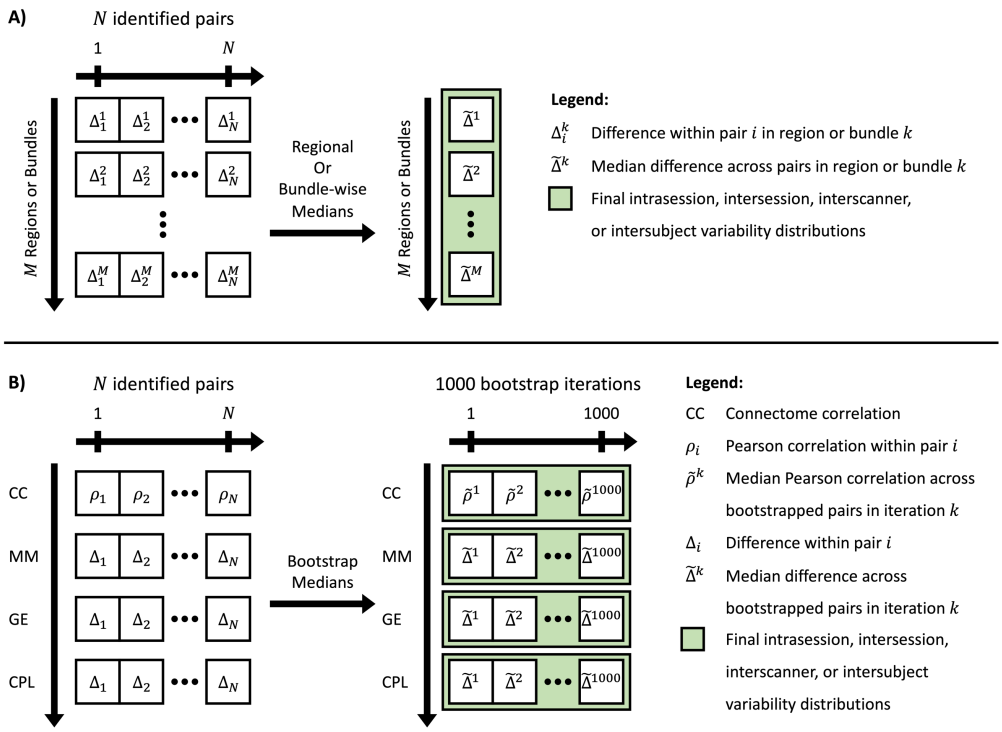

Figure 5. Summarizing variability at a given level of variation. (A) DTI, NODDI, and bundle segmentation. After differences within all N image pairs for all M regions or bundles are computed, the regional or bundle-wise medians are taken as the final distribution of variability for a given level of variation. (B) For the connectomics analysis, multiple regions or bundles are not used. Instead, a bootstrapped distribution of medians for each metric is taken as the final distribution of variability.

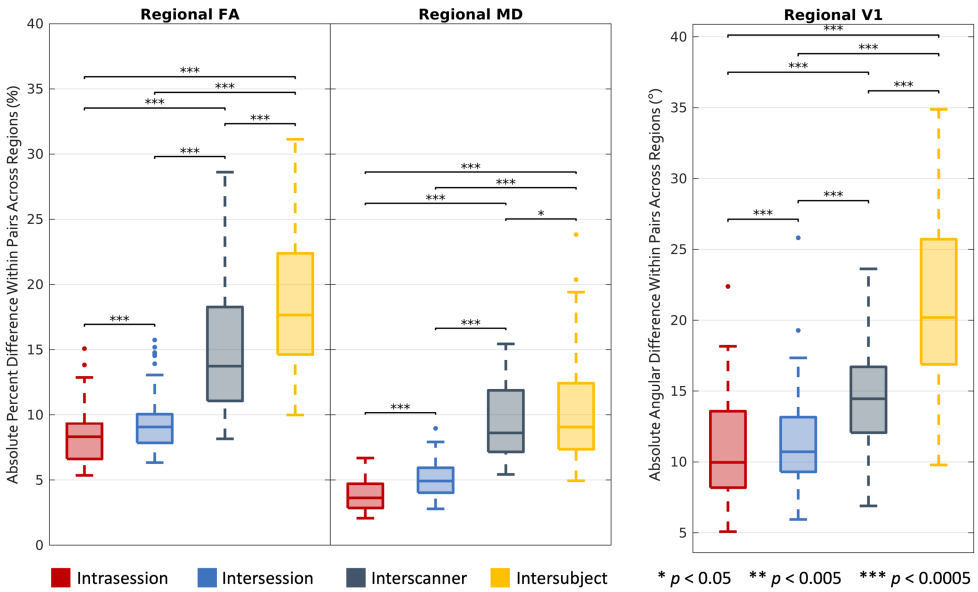

Figure 6. Variability in DTI. Visualization of differences within intrasession, intersession, interscanner, and intersubject pairs across 48 Johns Hopkins white matter atlas regions illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon signed-rank test.

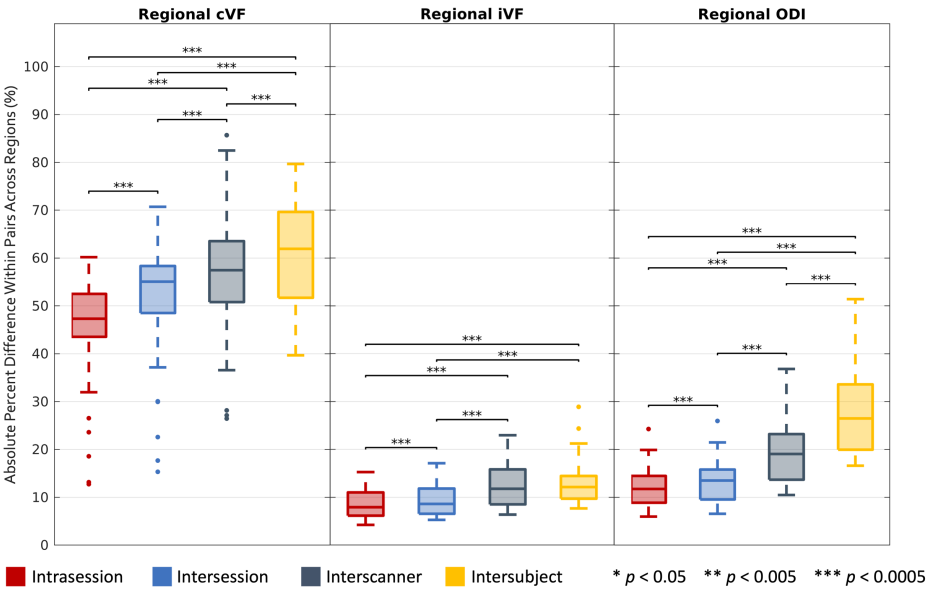

Figure 7. Variability in NODDI. Visualization of differences within intrasession, intersession, interscanner, and intersubject pairs across 48 Johns Hopkins white matter atlas regions consistently illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon signed-rank test.

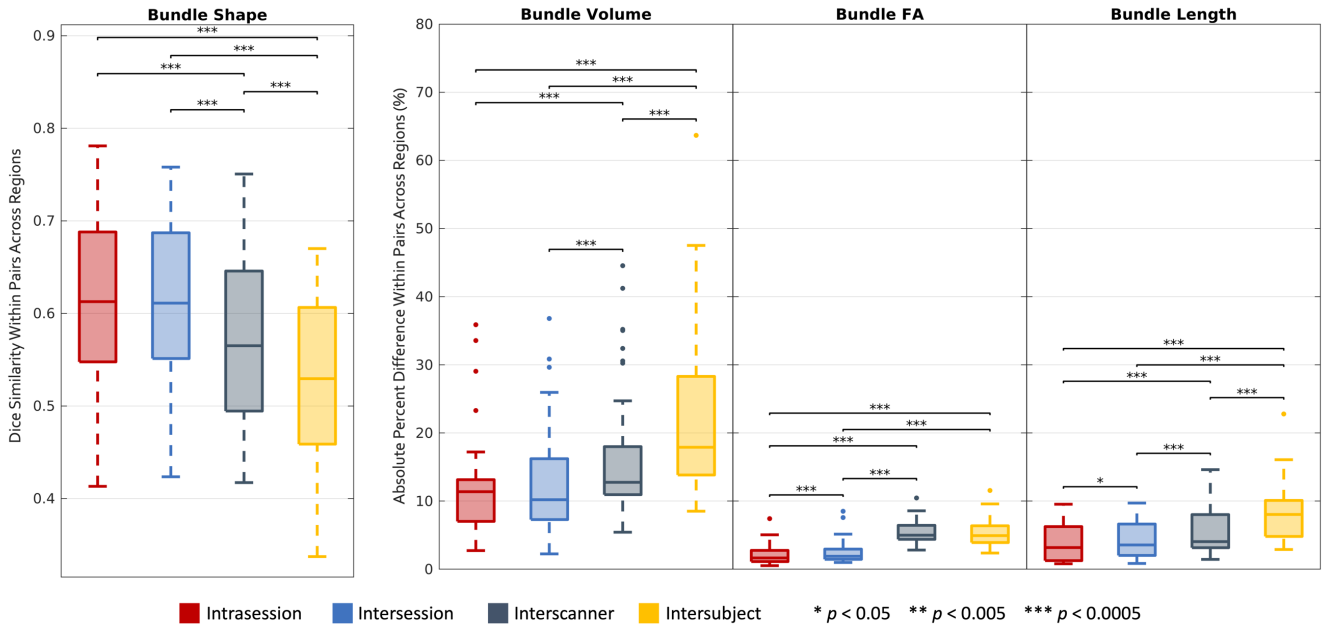

Figure 8. Variability in bundle segmentation. Visualization of differences within intrasession, intersession, interscanner, and intersubject pairs across 43 white matter bundles identified with the RecoBundles algorithm (Supporting Information Table S1) consistently illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon signed-rank test.

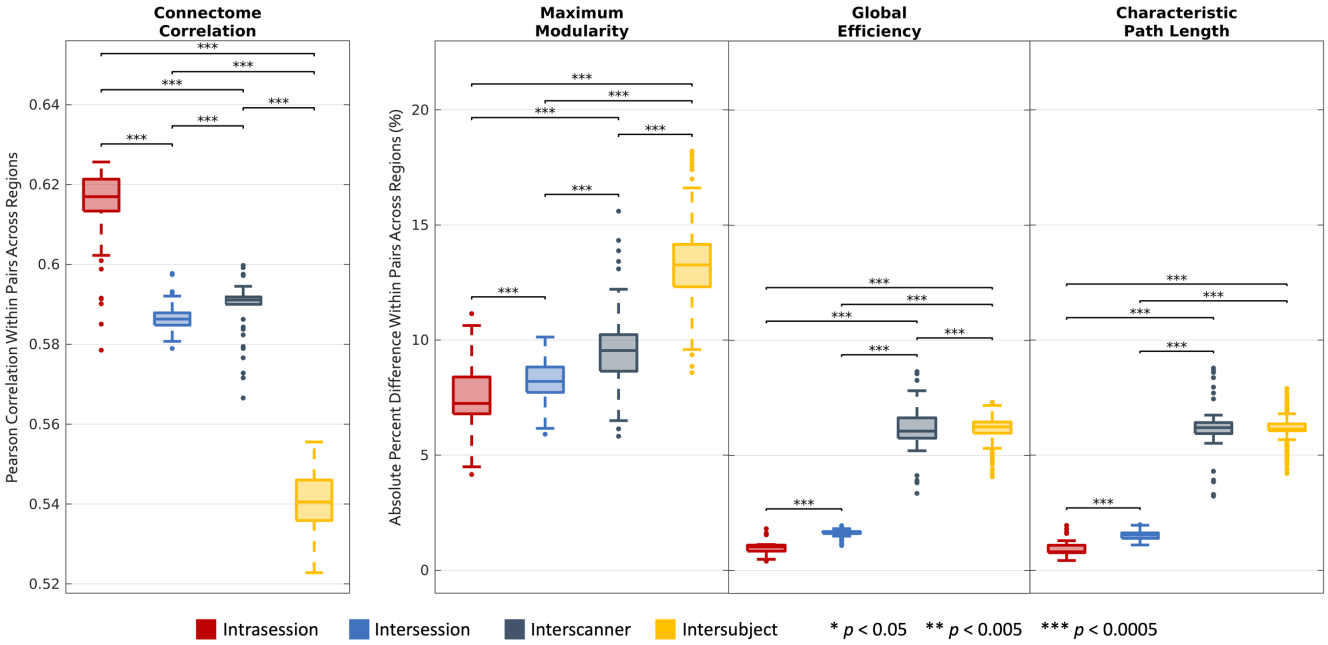

Figure 9. Variability in connectomics. With the exception of the intersession and interscanner correlation comparison, visualization of differences within intrasession, intersession, interscanner, and intersubject pairs consistently illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon rank-sum test.

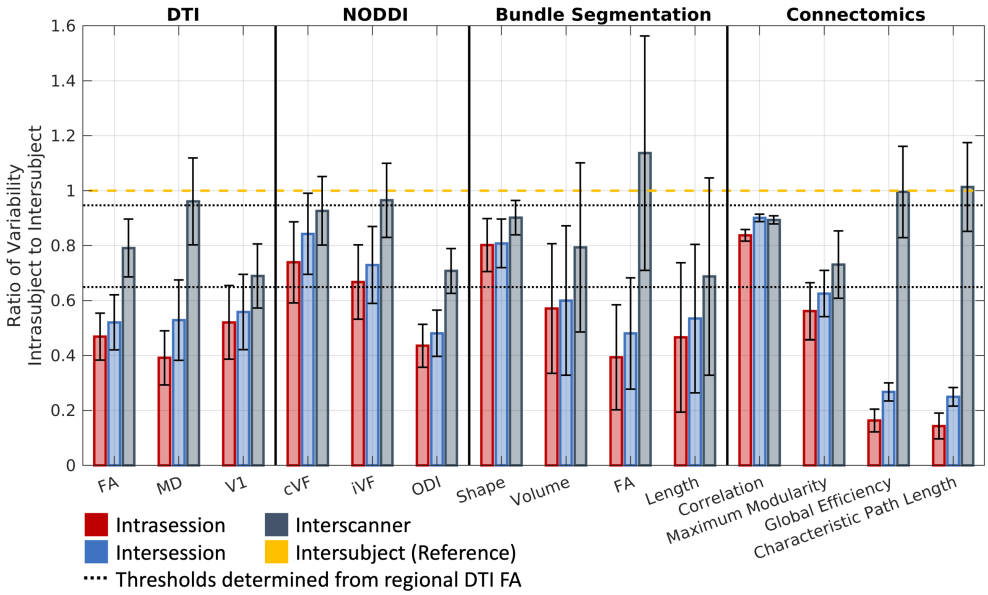

Figure 10. Comparing variability across processing approaches. The ratio of intrasubject to intersubject differences averaged across regions or bundles (DTI, NODDI, bundle segmentation) or bootstrapped distributions (connectomics) are shown with error bars denoting standard deviation. With the regional FA in the tensor-based model as a reference, the remaining DTI measurements as well as the NODDI, bundle segmentation, and connectomics measurements generally exhibit good (<0.65) to moderate (<0.95) intrasession and intersession variability, intrasubject variability that is non-negligible compared to intersubject variability, and interscanner effects that approach the size of intersubject effects.