

# MASiVar: Multisite, multiscanner, and multisubject acquisitions for studying variability in diffusion weighted MRI

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

<sup>1</sup>Department of Biomedical Engineering, Vanderbilt University, Nashville, Tennessee, USA

<sup>2</sup>Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA

<sup>3</sup>Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, Tennessee, USA

<sup>4</sup>Vanderbilt University Institute of Imaging Science, Vanderbilt University, Nashville, Tennessee, USA

<sup>5</sup>Department of Radiology, Mayo Clinic, Rochester, Minnesota, USA

<sup>6</sup>Department of Biomedical Engineering, The University of Texas at Austin, Austin, Texas, USA

<sup>7</sup>Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, New Jersey, USA

<sup>8</sup>Neuroscience Graduate Program, Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA

<sup>9</sup>Department of Psychology and Human Development, Peabody College, Vanderbilt University, Nashville, Tennessee, USA

<sup>10</sup>Department of Neurological Surgery, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

<sup>11</sup>Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

<sup>12</sup>Department of Intelligent Systems Engineering, Indiana University, Bloomington, Indiana, USA

<sup>13</sup>Department of Computer Science, Université de Sherbrooke, Sherbrooke, Québec, Canada

## Correspondence

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

## Funding information

National Institutes of Health (5R01EB017230, 5T32EB001628, 5T32GM007347, and 1UL1RR024975) and the National Science Foundation (1452485, 1660816, and 1750213)

**Purpose:** Diffusion-weighted imaging allows investigators to identify structural, microstructural, and connectivity-based differences between subjects, but variability due to session and scanner biases is a challenge.

**Methods:** To investigate DWI variability, we present MASiVar, a multisite data set consisting of 319 diffusion scans acquired at 3 T from  $b = 1000$  to  $3000 \text{ s/mm}^2$  across 14 healthy adults, 83 healthy children (5 to 8 years), three sites, and four scanners as a publicly available, preprocessed, and de-identified data set. 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, mean diffusivity, and principal eigenvector; region-wise CSF volume fraction, intracellular volume fraction, and orientation dispersion index; bundle-wise shape,

volume, fractional anisotropy, and length; and whole 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 it consistently 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 more globally investigate DWI variability across multiple levels and processing approaches simultaneously and suggests harmonization between scanners for multisite analyses should be considered before inference of group differences on subjects.

#### KEYWORDS

bundle segmentation, connectome, DTI, NODDI, reproducibility, variability

## 1 | INTRODUCTION

Diffusion-weighted MRI imaging is a noninvasive way of elucidating the brain's microstructural makeup.<sup>1</sup> Common modes of DWI analysis include representing the diffusion signal with tensors,<sup>2,3</sup> representing biological tissues with multi-compartment models,<sup>4-6</sup> identifying white-matter bundles,<sup>7</sup> and investigating the human structural connectome.<sup>8</sup> These approaches form the basis for many studies, including those investigating a wide range of neurological disorders including autism,<sup>9,10</sup> diabetes,<sup>11,12</sup> multiple sclerosis,<sup>13</sup> and schizophrenia,<sup>14</sup> as well as differences due to aging<sup>15</sup> and sex.<sup>16</sup> 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-matching 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 session and scanner effects relative to subject effect size, multiple scanned subjects are needed as well.

Another reason for this is the low number of properly configured publicly available data sets. Some of the few that exist that allow for investigations of DWI variability

are the MASSIVE (multiple acquisitions for standardization of structural imaging validation and evaluation) data set,<sup>17</sup> the Human Connectome Project (HCP) 3T data set,<sup>18</sup> the MICRA (microstructural image compilation with repeated acquisitions) data set,<sup>19</sup> the (SIMON) single individual volunteer for multiple observations across networks data set,<sup>20</sup> and the multisite data set published by Tong et al.<sup>21</sup> The MASSIVE data set consists of 1 subject scanned repeatedly on one scanner<sup>17</sup>; the HCP data set consists of multiple subjects with multiple acquisitions per session all on one scanner<sup>18</sup>; the MICRA data set consists of multiple subjects scanned repeatedly on one scanner<sup>19</sup>; the SIMON data set consists of 1 subject scanned at over 70 sites<sup>20</sup>; and the Tong et al. data set consists of multiple subjects each scanned on multiple scanners.<sup>21</sup>

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 data sets 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 data set able to characterize DWI variability due to session, scanner, and subject effects.

TABLE 1 Survey of existing DWI variability estimates against those presented in the present work.

Approach	Measure	Intrasession	Intersession	Interscanner	Intersubject	Citation		
DTI	FA	<b>3.34% CoV</b>	<b>5.29% CoV</b>	<b>8.78% CoV</b>	<b>11.95% CoV</b>	<b>Present work</b>		
		2% CoV	3% CoV	–	–	(Farrell 2010) <sup>33</sup>		
		–	1% CoV	3% CoV	–	(Magnotta 2012) <sup>50</sup>		
		–	1-2% CoV	2-4% CoV	–	(Vollmar 2010) <sup>51</sup>		
		–	3% CoV	8% CoV	–	(Palacios 2017) <sup>52</sup>		
		–	0.5% CoV	2% CoV	–	(Andica 2020) <sup>53</sup>		
		–	0.90-0.99 ICC	0.82-0.99 ICC	–	(Vollmar 2010) <sup>51</sup>		
		–	0.74-1.00 ICC	0.54-0.97 ICC	–	(Andica 2020) <sup>53</sup>		
		–	0.6%-1% CoV	–	–	(Koller 2020) <sup>19</sup>		
		–	0.93-0.97 ICC	–	–	(Koller 2020) <sup>19</sup>		
		–	~0.95 PC	–	–	(Koller 2020) <sup>19</sup>		
		–	2.1 % CoV	2.0% CoV	3.8% CoV	(Grech-Sollars 2015) <sup>75</sup>		
		–	0.53 ICC	0.47 ICC	–	(Grech-Sollars 2015) <sup>75</sup>		
			MD	<b>1.37% CoV</b>	<b>3.43% CoV</b>	<b>6.22% CoV</b>	<b>5.12% CoV</b>	<b>Present work</b>
				1% CoV	1% CoV	–	–	(Farrell 2010) <sup>33</sup>
–	1% CoV			2% CoV	–	(Magnotta 2012) <sup>50</sup>		
–	2% CoV			6% CoV	–	(Palacios 2017) <sup>52</sup>		
–	0.2% CoV			3% CoV	–	(Andica 2020) <sup>53</sup>		
–	0.5%-1% CoV			–	–	(Koller 2020) <sup>19</sup>		
–	0.94-0.96 ICC			–	–	(Koller 2020) <sup>19</sup>		
–	~0.66 PC			–	–	(Koller 2020) <sup>19</sup>		
–	1.3% CoV			1.6% CoV	2.6% CoV	(Grech-Sollars 2015) <sup>75</sup>		
–	0.41 ICC			0.59 ICC	–	(Grech-Sollars 2015) <sup>75</sup>		
	V1			<b>4.49° AV</b>	<b>7.28° AV</b>	<b>9.48° AV</b>	<b>13.42° AV</b>	<b>Present work</b>
				~2°-8° AV	~7°-12° MAD	–	–	(Farrell 2010) <sup>33</sup>
NODDI	cVF			<b>27.33% CoV</b>	<b>34.57% CoV</b>	<b>40.34% CoV</b>	<b>53.11% CoV</b>	<b>Present work</b>
				–	1.6%-3.6% CoV	15.9% CoV	–	(Andica 2020) <sup>53</sup>
		–	0.133-0.997 ICC	0.013-0.545 ICC	–	(Andica 2020) <sup>53</sup>		
		iVF	<b>3.64% CoV</b>	<b>5.48% CoV</b>	<b>7.89% CoV</b>	<b>8.27% CoV</b>	<b>Present work</b>	
			–	0.4% CoV	0.9% CoV	–	(Andica 2020) <sup>53</sup>	
			–	0.773-0.989 ICC	0.300-0.935 ICC	–	(Andica 2020) <sup>53</sup>	
	ODI	<b>4.56% CoV</b>	<b>6.49% CoV</b>	<b>13.14% CoV</b>	<b>19.54% CoV</b>	<b>Present work</b>		
		–	0.2%-0.3% CoV	4.2% CoV	–	(Andica 2020) <sup>53</sup>		
		–	0.789-0.998 ICC	0.181-0.962 ICC	–	(Andica 2020) <sup>53</sup>		
	–	5.7% CoV	–	–	(Tariq 2013) <sup>54</sup>			

(Continues)

TABLE 1 (Continued)

Approach	Measure	Intraseession	Intersession	Interscanner	Intersubject	Citation
Bundle segmentation	Shape	<b>0.82 DV</b>	<b>0.81 DV</b>	<b>0.76 DV</b>	<b>0.68 DV</b>	<b>Present work</b>
		~0.67 Dice	~0.64 Dice	~0.58 Dice	–	(Nath 2020) <sup>56</sup>
		–	0.65-0.92 Dice	–	–	(Besseling 2012) <sup>55</sup>
		–	0.72 wDice	–	–	(Cousineau 2017) <sup>76</sup>
		–	0.71-0.87 wDice	–	–	(Boukadi 2019) <sup>77</sup>
	–	–	–	–	~0.5-0.6 Dice	(Schilling 2020) <sup>78</sup>
	Volume	<b>4.63% CoV</b>	<b>5.82% CoV</b>	<b>9.07% CoV</b>	<b>15.04% CoV</b>	<b>Present work</b>
		–	3%-22% CoV	–	–	(Besseling 2012) <sup>55</sup>
		–	0.53-0.96 ICC	–	–	(Besseling 2012) <sup>55</sup>
		–	0.41-0.83 ICC	–	–	(Boukadi 2019) <sup>77</sup>
	FA	<b>0.71% CoV</b>	<b>1.10% CoV</b>	<b>3.15% CoV</b>	<b>2.47% CoV</b>	<b>Present work</b>
		–	1%-4% CoV	–	–	(Besseling 2012) <sup>55</sup>
		–	0.65-0.94 ICC	–	–	(Besseling 2012) <sup>55</sup>
		–	0.62-0.89 ICC	–	–	(Boukadi, 2019) <sup>77</sup>
Length	<b>1.27% CoV</b>	<b>1.93% CoV</b>	<b>2.42% CoV</b>	<b>6.11% CoV</b>	<b>Present work</b>	
	–	0.68-0.89 ICC	–	–	(Boukadi, 2019) <sup>77</sup>	
Connectomics	Whole connectome	<b>0.89 PCV</b>	<b>0.89 PCV</b>	<b>0.85 PCV</b>	<b>0.80 PCV</b>	<b>Present work</b>
		–	0.6-0.95 PC	–	–	(Prčkovska 2016) <sup>57</sup>
	–	–	–	–	32.7%-39.9% CD	(Girard 2015) <sup>79</sup>
	MM	<b>3.29% CoV</b>	<b>3.83% CoV</b>	<b>6.49% CoV</b>	<b>14.55% CoV</b>	<b>Present work</b>
	GE	<b>0.44% CoV</b>	<b>0.91% CoV</b>	<b>3.38% CoV</b>	<b>3.80% CoV</b>	<b>Present work</b>
	–	31% CoV	–	–	(Roine 2019) <sup>58</sup>	
	–	0.78 ICC	–	–	(Roine 2019) <sup>58</sup>	
	CPL	<b>0.40% CoV</b>	<b>0.93% CoV</b>	<b>3.52% CoV</b>	<b>3.76% CoV</b>	<b>Present work</b>
–	2% CoV	–	–	(Roine 2019) <sup>58</sup>		
–	0.77 ICC	–	–	(Roine 2019) <sup>58</sup>		

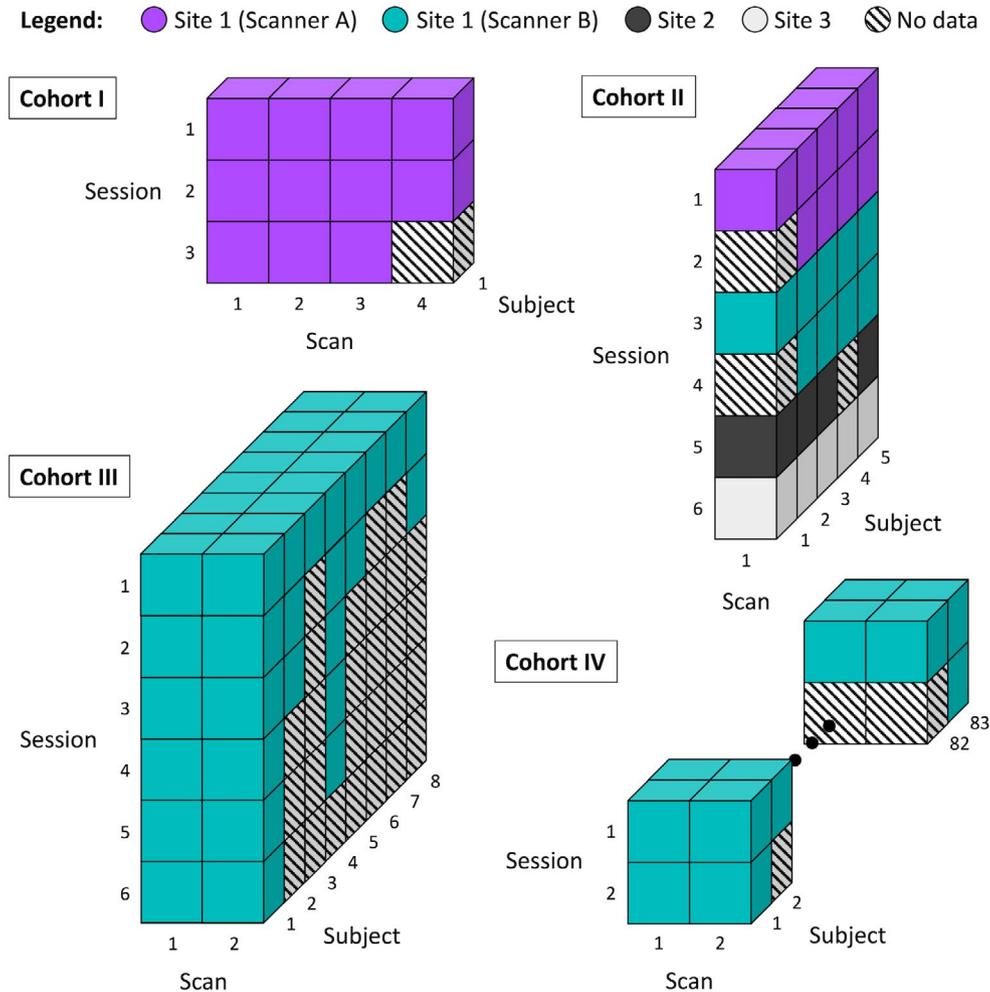
Abbreviations: AV, angular variation; CD, connectome distance; CoV, coefficient of variation; CPL, characteristic path length; cVF, CSF volume fraction; DV, Dice variation; FA, fractional anisotropy; GE, global efficiency; ICC, intraclass correlation coefficient; iVF, intracellular volume fraction; MAD, mean angular difference; MD, mean diffusivity; MM, maximum modularity; ODI, orientation dispersion index; PC, Pearson correlation; PCV, Pearson correlation variation; VI, principal eigenvector; wDice, weighted Dice; –, not investigated.

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.

## 2 | METHODS

### 2.1 | Data acquisition

The MASiVar data set consists of data acquired from 2016 to 2020 to study both DWI variability and other phenomena.



**FIGURE 1** Overview of the MASiVar data set. This data set consists of four cohorts. Cohort I consists of 1 adult subject scanned repeatedly on one scanner. This subject underwent three separate imaging sessions and acquired three to four scans per session. Cohort II consists of 5 adult subjects each scanned on three to four different scanners across three institutions. Each subject underwent one to two 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 one to six 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 one to two sessions on the scanner and had two scans acquired per session.

As such, the data exist in four cohorts, designated as I, II, III, and IV (Figure 1).

Cohort I consists of 1 healthy adult subject (male, age 25 years) with multiple imaging sessions on a 3T Philips Achieva scanner (Amsterdam, the Netherlands) 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  $3000 \text{ s/mm}^2$  (Table 2). These scans were acquired at 2.5-mm isotropic resolution with  $\text{TE/TR} = 94 \text{ ms}/2650 \text{ ms}$ .

Cohort II consists of 5 healthy adult subjects (3 male, 2 female, age 27–47 years) scanned for one to two sessions on each of three to four different scanners. Each subject underwent all sessions within 1 year. The scanners included scanner A, another 3T Philips Achieva scanner at site 1 (scanner B), a 3T General Electric Discovery MR750 scanner (Boston, MA) at

site 2, and a 3T Siemens Skyra scanner (Erlangen, Germany) 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)  $\text{s/mm}^2$  and a 30- or 32-direction acquisition at  $b = 1000 \text{ s/mm}^2$  (Table 2). The scans acquired on scanner B, at site 2, and at site 3, and all the 30-direction or 32-direction scans were acquired at 2.5-mm isotropic resolution. On scanner A, one subject's 96-direction acquisitions were also acquired at 2.5-mm isotropic resolution, while the remainder were acquired at  $1.9 \times 1.9 \times 2.2 \text{ mm}$  (sagittal, coronal, and axial) resolution. For acquisitions on scanner A, the 2.5-mm isotropic 96-direction scans were acquired with  $\text{TE/TR} = 90 \text{ ms}/5200 \text{ ms}$ , whereas the other 96-direction acquisitions were acquired with  $\text{TE/TR} = 90 \text{ ms}/5950 \text{ ms}$ , and  $\text{TE/TR} = 55 \text{ ms}/6127 \text{ ms}$  to  $7309 \text{ ms}$  for the 32-direction acquisitions. For acquisitions on scanner B, the 96-direction scans were

**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

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

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

Cohort IV consists of 83 healthy child subjects (48 male, 35 female, ages 5-8 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 1 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<sup>2</sup> and a 56-direction  $b = 2000$  s/mm<sup>2</sup> acquisition (Table 2). These scans were acquired at  $2.1 \times 2.1 \times 2.2$  mm (sagittal, coronal, and axial) resolution with TE/TR = 79 ms/2900 ms.

All acquisitions were phase-encoded in the posterior–anterior direction and were acquired with one  $b = 0$  s/mm<sup>2</sup> volume each. Reverse phase-encoded ( $b = 0$  s/mm<sup>2</sup> volumes were also acquired for all scans in all cohorts except for those from 1 subject in cohort II at site 3. Most sessions also included a T<sub>1</sub>-weighted image for structural analysis or distortion correction.<sup>22</sup> All images were de-identified and all scans were acquired only after informed consent under supervision of the project institutional review board.

## 2.2 | Data preprocessing

After acquisition, all scans in MASiVar were preprocessed and quality checked with the PreQual pipeline.<sup>23</sup> In brief, all acquisitions per scan were denoised with the Marchenko-Pastur technique,<sup>24–26</sup> intensity normalized, and distortion corrected. Distortion correction included susceptibility-induced distortion correction<sup>27</sup> using reverse phase-encoded  $b = 0$  s/mm<sup>2</sup> volumes when available and the Synb0-DisCo deep learning framework<sup>22</sup> and associated T<sub>1</sub> image when not, eddy current-induced distortion correction, intervolume motion correction, and slice-wise signal dropout imputation.<sup>28,29</sup> The estimated volume-to-volume displacement corrected during preprocessing, and SNRs of the scans are reported in Supporting Information Figure S1.

## 2.3 | 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 DTI signal representation;
2. A multicompartment neurite orientation dispersion and density imaging (NODDI) model<sup>4</sup>;
3. The RecoBundles white-matter bundle segmentation technique<sup>30</sup>; and
4. A connectomics representation with graph-based measures.<sup>31</sup>

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 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.

## 2.4 | Defining intrasession, intersession, interscanner, and intersubject groups

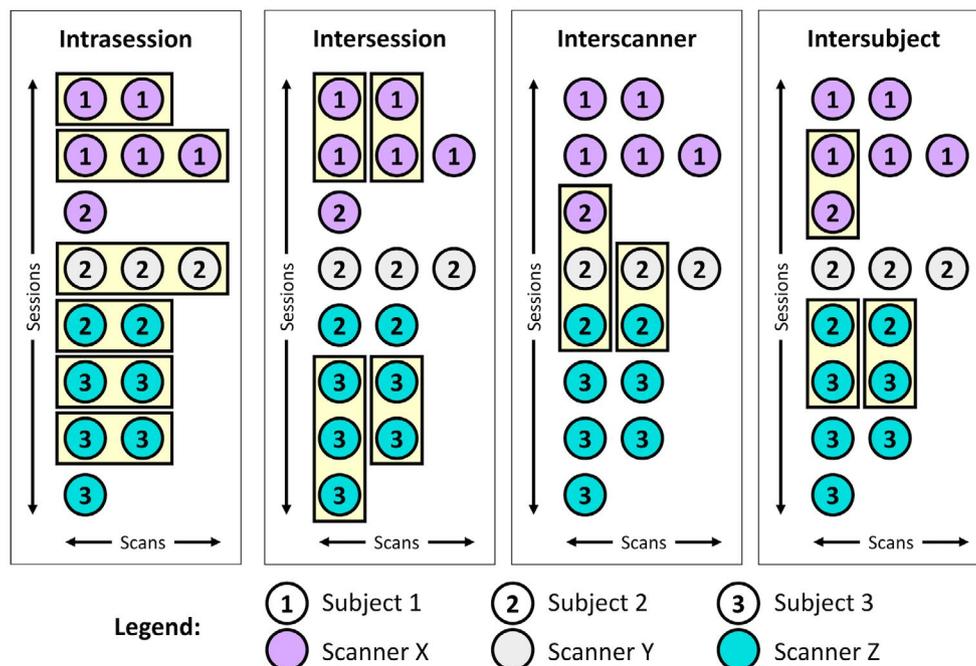
To investigate variability, we first identify qualifying “groups” of intrasession, intersession, interscanner, and intersubject scans from cohorts I to III in MASiVar (Figure 2). We define an intrasession group as any session with at least two scans. Because sessions are necessarily nested in scanners and subjects, these samples are distributed across scanners and subjects. We find 24 qualifying groups, each containing two to four scans. To form an intersession group, we randomly select one scan from each of a subject’s different sessions on the same scanner. We repeat this process without replacement to form additional groups until no more groups with at least two scans can be formed. We find 22

qualifying groups, each containing two to six scans. As with the intrasession groups, these groups are distributed across scanners and subjects. To form an interscanner group, we randomly select one scan from each of a subject’s sessions on different scanners and repeat this process without replacement to form additional groups until no more groups with at least two scans can be formed. These groups are distributed across subjects. We find nine groups, each containing two to four scans. To form an intersubject group, we randomly select one scan from each of the different subjects scanned on one scanner and repeat this process without replacement to form additional groups until no more groups with at least two scans can be formed. We find 14 qualifying groups, each containing 2-13 scans, distributed across the four scanners used in MASiVar.

## 2.5 | Computing variability

Overall, we evaluate the variability for a given effect by first computing variability within each group and then visualizing the distribution across groups on the intrasession, intersession, interscanner, and intersubject levels. To compare across levels, we use six pair-wise non-parametric Wilcoxon rank-sum statistical tests with an uncorrected significance level of 0.05 and a Bonferroni-corrected significance level of 0.008.<sup>32</sup>

We compute variability with the coefficient of variation (CoV) for scalar metrics, angular variation (AV) for V1,



**FIGURE 2** Example identification of scan groups at the four levels of variation. The MASiVar data set consists of scans across multiple sessions, scanners, and subjects that can be grouped in order to satisfy intrasession, intersession, interscanner, and intersubject criteria. The scans in each of these groups should produce the same measurements; thus, quantification of variation within groups provides an estimate of variability. For the intersession, interscanner, and intersubject groups, scans are randomly shuffled within sessions before grouping.

Dice variation (DV) for bundle shape, and Pearson correlation variation (PCV) for whole connectome variability. These variability metrics are mathematically defined as follows (Equations 1-4), and their uses are further refined for the different DWI approaches in the following sections.

CoV (%) is defined for each group as the SD of the scalar metrics in each group,  $\sigma$ , divided by the mean of the group,  $\bar{x}$ , times 100% (Equation 1). Intuitively, CoV is computed as the proportion of the average scalar measurement attributable to variability. As such, as variability increases, so does CoV.

$$\text{CoV}(\%) = 100\% \times \frac{\sigma}{\bar{x}} \quad (1)$$

AV ( $^{\circ}$ ) is defined for each group as the average angle between the  $N$  members of the group, defined with unit vectors, and the group average unit vector,  $\bar{x}$  (Equation 2).<sup>33</sup> As principal eigenvectors are direction agnostic,  $\bar{x}$  is computed iteratively to ensure the vectors are oriented correctly. We (1) compute  $\bar{x}$ , (2) identify all vectors oriented  $>90^{\circ}$  from  $\bar{x}$ , (3) negate those vectors, and (4) repeat steps 1-3 until step 2 identifies no additional vectors. The AV is computed on the reoriented vectors as follows (Equation 2). Intuitively, AV is computed as the radius of the cone of uncertainty around the average eigenvector measurement in degrees. As variability increases, so does AV.

$$\text{AV}(\circ) = \frac{1}{N} \sum_{i=1}^N \cos^{-1}(|x_i \cdot \bar{x}|) \quad (2)$$

DV (range: 0-1) is defined for each group as the average Dice similarity coefficient,  $DSC$ , between the  $N$  bundles in the group, represented with binary masks, and the group average bundle,  $\bar{x}$  (Equation 3).<sup>34</sup> The value of  $\bar{x}$  is computed with a voxel-wise majority vote. Intuitively, DV extends the concept of “the cone of uncertainty” described for AV to higher dimensions around the average segmented bundle. However, unlike AV that describes how “large” the radius is with a distance metric, DV describes how “small” it is with the Dice similarity metric. As such, as variability increases, DV decreases.

$$\text{DV} = \frac{1}{N} \sum_{i=1}^N DSC(x_i, \bar{x}) \quad (3)$$

PCV (range: -1 to 1) is defined for each group as the average Pearson correlation,  $\rho$ , between the  $N$  connectomes of the group and the group average connectome,  $\bar{x}$  (Equation 4). Similar to DV, PCV is computed as the radius of the extended “cone of uncertainty” around the average connectome with the Pearson correlation similarity metric. Thus, as variability increases, PCV decreases.

$$\text{PCV} = \frac{1}{N} \sum_{i=1}^N \rho(x_i, \bar{x}) \quad (4)$$

## 2.6 | Variability in DTI and NODDI

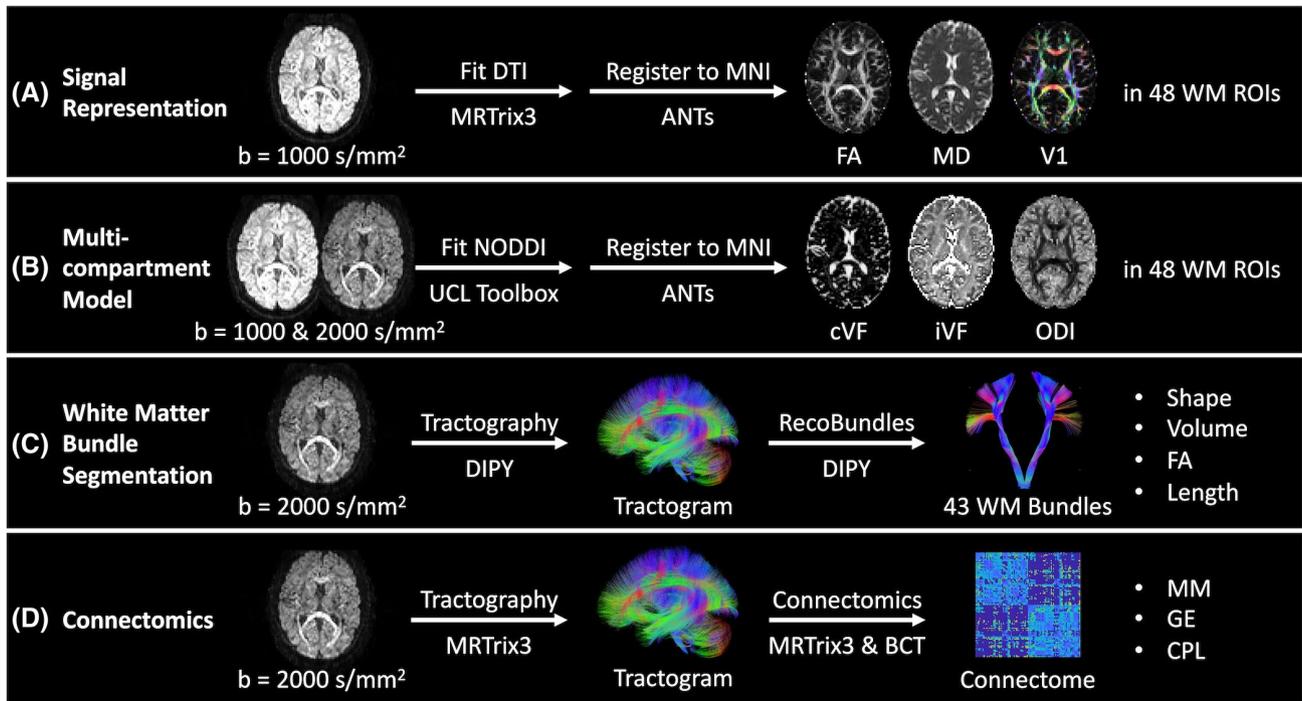
For the DTI approach, we extract the  $b = 1000 \text{ s/mm}^2$  acquisition from each scan with the largest number of directions. We then calculate the diffusion tensor for each scan using an iteratively reweighted least squares approach implemented in *MRtrix3*.<sup>35</sup> The tensors are subsequently converted to FA, MD, and V1 representations of the data.<sup>36</sup> These images are then deformably registered to the Montreal Neurological Institute (MNI) image space with the *ANTs* software package.<sup>37,38</sup> From there, we identify the 48 regions of interest in each image defined by the Johns Hopkins white-matter atlas<sup>39-41</sup> (Figure 3A).

For the NODDI approach, we extract the  $b = 1000 \text{ s/mm}^2$  acquisition from each scan with the largest number of directions and the  $b = 2000 \text{ s/mm}^2$  acquisition. We then fit the multicompartiment model with the University College London NODDI Toolbox as implemented in *MATLAB* (Natick, MA).<sup>4</sup> 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 regions of interest in each image defined by the Johns Hopkins white-matter atlas (Figure 3B).

We perform the DTI and NODDI variability calculations on a regional basis in MNI space with voxel-wise correspondence between images. For FA, MD, cVF, iVF, and ODI, we compute the CoV for each region as the median voxel-wise CoV. We report the regional median across the groups for each level. Similarly, for V1 we compute the AV for each region as the median voxel-wise AV and report the regional median across the groups for each level.

## 2.7 | Variability in bundle segmentation

For the white-matter segmentation approach, we extract the  $b = 2000 \text{ s/mm}^2$  acquisition from each scan. We calculate a whole-brain tractogram with *DIPY* of 2 million streamlines.<sup>42</sup> We use the constrained spherical deconvolution model<sup>43</sup> with probabilistic local tracking with a maximum angle of  $25^{\circ}$ , a seeding criterion of  $\text{FA} > 0.3$ , and a stopping criterion of  $\text{FA} < 0.2$ . 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.<sup>30</sup> 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



**FIGURE 3** Outline of processing and measurements investigated presently in four common diffusion MRI analysis approaches. A,B, We quantify variability in the tensor-based fractional anisotropy (FA), mean diffusivity (MD), and principal eigenvector (V1) measurements and neurite orientation dispersion and density imaging (NODDI)-based CSF volume fraction (cVF), intracellular volume fraction (iVF), and orientation dispersion index (ODI) measurements in Montreal Neurological Institute (MNI) space in 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 maximum modularity (MM), global efficiency (GE), and characteristic path length (CPL) scalar graph measures.

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<sup>35</sup> and then to an FA representation.<sup>36</sup> Each bundle's binary mask is then applied to obtain the median voxel-wise FA value per bundle (Figure 3C).

Unlike the DTI and NODDI cases, streamline-wise and subsequent voxel-wise correspondence cannot be achieved with tractography and bundle segmentation, so we compute variability on a bundle-wise basis. For bundle shape, we compute the DV on the binary masks for each bundle, and for volume, FA, and length we compute the CoV for each bundle. We report the bundle-wise median across the groups for each level for each of these measures.

## 2.8 | Variability in connectomics

For the connectomics approach, we extract the  $b = 2000$   $\text{s/mm}^2$  acquisition from each scan. We then calculate a whole-brain tractogram with *MRtrix3*.<sup>44</sup> We first use the constrained spherical deconvolution model with probabilistic tracking with a maximum angle of  $25^\circ$ , a seeding

criterion of  $\text{FA} > 0.3$ , and a stopping criterion of  $\text{FA} < 0.2$  to calculate a 10 million streamline tractogram. The tractogram is then filtered with the SIFT approach to 2 million streamlines.<sup>45</sup> We parcellate the brain into 96 cortical regions using the Harvard-Oxford cortical atlas<sup>46–49</sup> 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*<sup>31</sup> (Figure 3D).

To evaluate whole connectome variability, we report the PCV across the groups for each level. To evaluate variability in the MM, GE, and CPL graph measures, we report the CoV across the groups for each level.

## 2.9 | Comparing variability across processing approaches

Last, to obtain a more global understanding of the session, scanner, and subject effects across the four different processing approaches, we compare the median CoV estimates for FA and MD (DTI), cVF, iVF, and ODI (NODDI), volume, FA, and length (bundle segmentation), and MM, GE, and

CPL (connectomics) on the intrasession, intersession, interscanner, and intersubject levels. We determine differences with six pair-wise Wilcoxon signed-rank tests at an uncorrected significance level of 0.05 and a Bonferroni-corrected significance of 0.008.

### 3 | RESULTS

#### 3.1 | Variability in DTI

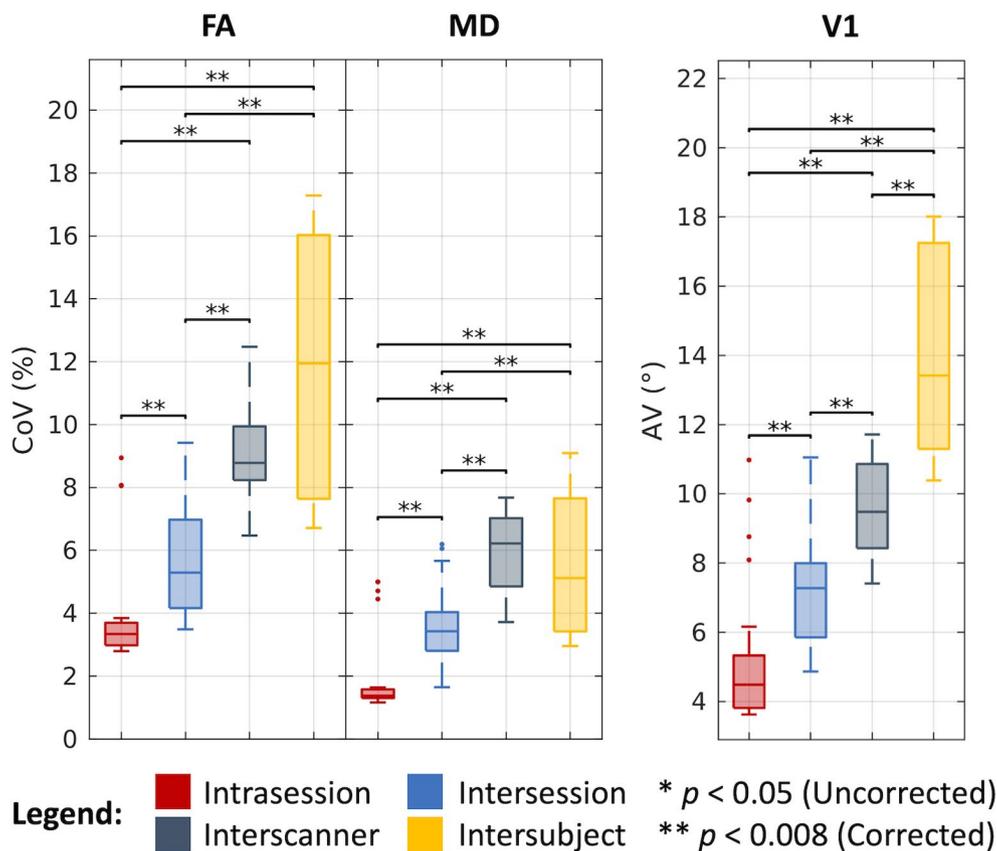
As shown in Figure 4 and tabulated in Table 1, we find that the median CoV for FA across intrasession groups is 3.34%, across intersession groups is 5.29%, across interscanner groups is 8.78%, and across intersubject groups is 11.95%. We find the corresponding estimates in the MD case to be 1.37%, 3.43%, 6.22%, and 5.12%, and the corresponding AV estimates in the V1 case to be 4.49°, 7.28°, 9.48°, and 13.42°, respectively. The differences between most of these estimates are statistically significant after Bonferroni correction ( $P < .008$ ; Wilcoxon rank-sum test). Notably, we find for the FA and MD cases that interscanner variability is comparable to intersubject variability.

#### 3.2 | Variability in NODDI

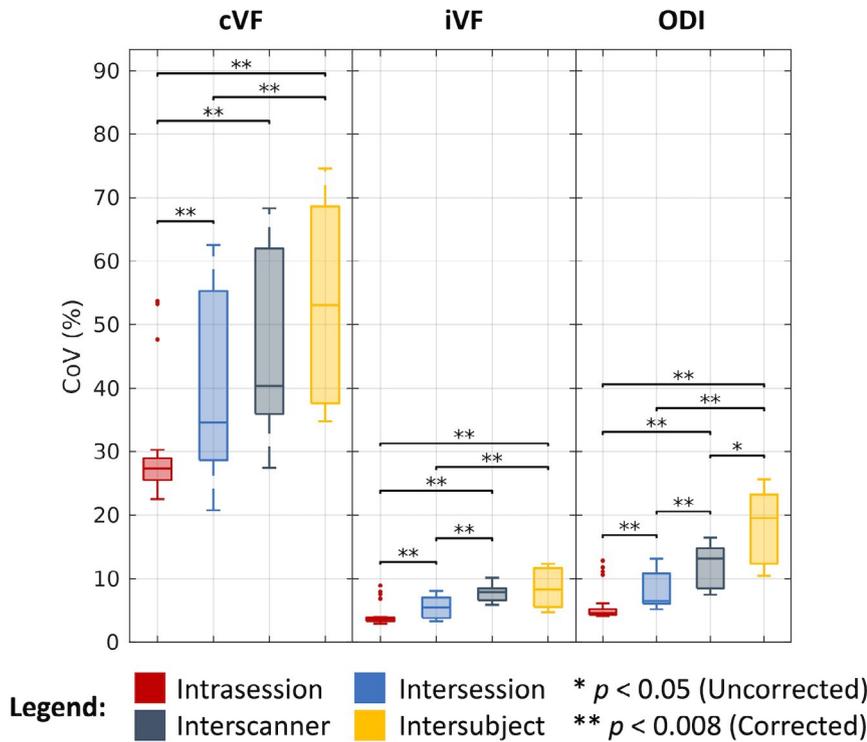
As shown in Figure 5 and tabulated in Table 1, we find that the median CoV for cVF across intrasession groups is 27.33%, across intersession groups is 34.57%, across interscanner groups is 40.34%, and across intersubject groups is 53.11%. We find the corresponding estimates in the iVF case to be 3.64%, 5.48%, 7.89%, and 8.27%, and in the ODI case to be 4.56%, 6.49%, 13.14%, and 19.54%, respectively. As with the DTI case, most of these estimates are statistically different after Bonferroni correction ( $P < .008$ ; Wilcoxon rank-sum test). Of note, we evaluated cVF only in white-matter regions defined by the Johns Hopkins atlas and thus dealt with very low cVF values when computing CoV. Additionally, we find that for the cVF and iVF cases that interscanner variability is comparable to intersubject variability.

#### 3.3 | Variability in bundle segmentation

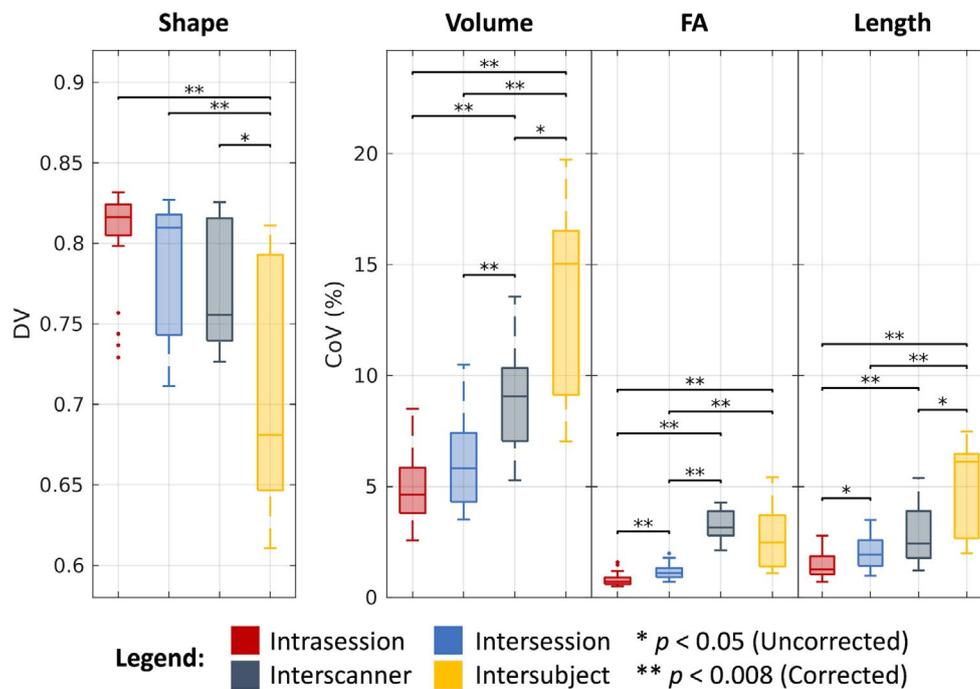
As shown in Figure 6 and tabulated in Table 1, we find that bundles overlap at a median DV of 0.82 across intrasession groups, 0.81 across intersession groups, 0.76 across interscanner groups, and 0.68 across intersubject groups.



**FIGURE 4** Variability in DTI. Visualization of variation across intrasession, intersession, interscanner, and intersubject groups illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon rank-sum test with and without Bonferroni correction.



**FIGURE 5** Variability in NODDI. Visualization of variation across intrasession, intersession, interscanner, and intersubject groups illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon rank-sum test with and without Bonferroni correction.



**FIGURE 6** Variability in bundle segmentation. Visualization of variation across intrasession, intersession, interscanner, and intersubject groups illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon rank-sum test with and without Bonferroni correction.

We find the median CoV estimates for the corresponding levels of variation across groups in the bundle volume case to be 4.63%, 5.82%, 9.07%, and 15.04%, in the FA case to be 0.71%, 1.10%, 3.15%, and 2.47%, and in the bundle length case to be 1.27%, 1.93%, 2.42%, and 6.11%, respectively. As

with the DTI and NODDI cases, most of these estimates are statistically different after Bonferroni correction ( $P < .008$ , Wilcoxon rank-sum test). Notably, we find that in the FA case, interscanner variability is comparable to intersubject variability.

### 3.4 | Variability in connectomics

As shown in Figure 7 and tabulated in Table 1, we find that the whole connectomes correlate at a median PCV of 0.89 across intrasession groups, 0.89 across intersession groups, 0.85 across interscanner groups, and 0.80 across intersubject groups. We find the median CoV estimates for the corresponding levels of variation across groups in the MM case to be 3.29%, 3.83%, 6.49%, and 14.55%, in the GE case to be 0.44%, 0.91%, 3.38%, and 3.80%, and in the CPL case to be 0.40%, 0.93%, 3.52%, and 3.76%, respectively. As with the other processing approaches, most of these estimates are statistically different after Bonferroni correction ( $P < .008$ ; Wilcoxon rank-sum test). Additionally, we note that for both the GE and CPL cases, interscanner variability is comparable to intersubject variability.

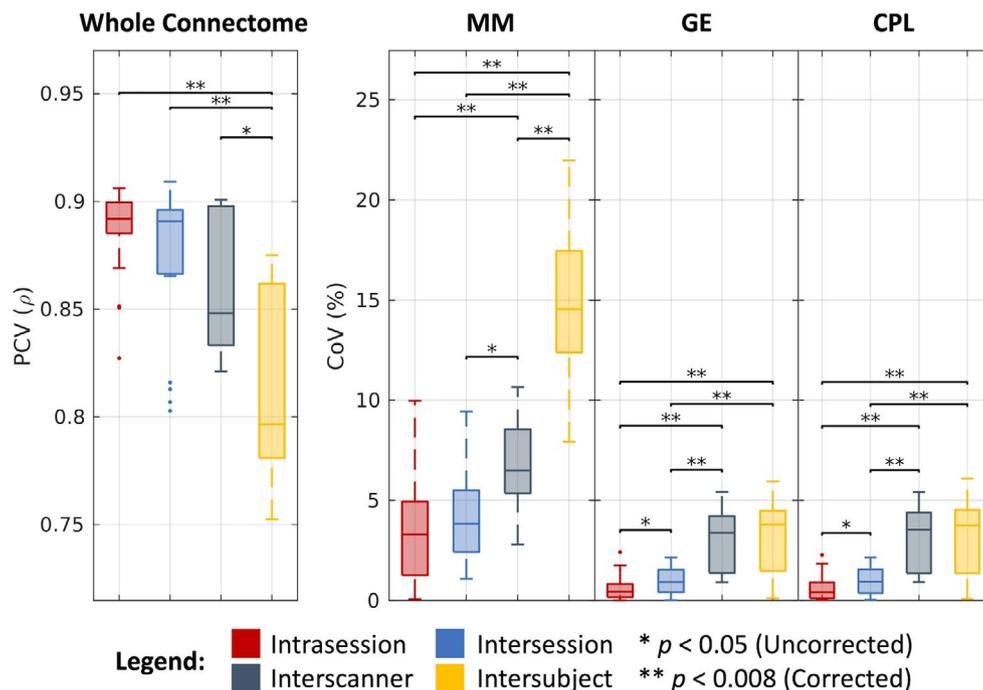
### 3.5 | Comparing variability across processing approaches

As shown in Figure 8, we find that the overall CoV estimates across the four processing approaches increase with consideration of intrasession, intersession, interscanner, and intersubject effects. Additionally, we find that all of these estimates are statistically different after Bonferroni correction, with the exception of the interscanner and intersubject comparison. Last, with the exception of the outlier (cVF in white

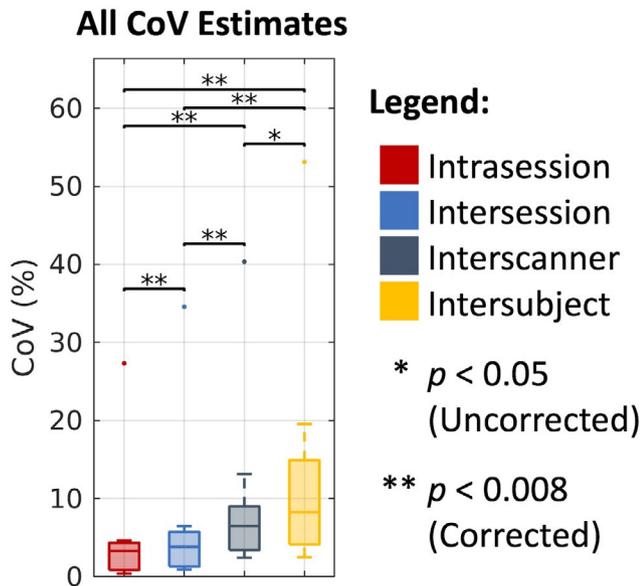
matter), we note that all of the approaches exhibit similar variability within each level, with a median CoV of 3.29% on the intrasession level, 3.83% on the intersession level, 6.49% on the interscanner level, and 8.27% on the intersubject level.

## 4 | DISCUSSION AND CONCLUSIONS

Here, we present MASiVar, a data set designed for investigation of DWI variability. Additionally, to demonstrate the capacity of MASiVar as a resource, we characterize intrasession, intersession, interscanner, and intersubject variability in four common DWI processing approaches. In support of our hypothesis, we consistently find that variability increases with consideration of session, scanner, and subject effects. We also find that overall and for each of the four approaches, interscanner variability can approach or even be comparable to intersubject variability. Last, we find that most of the DWI scalar measurements investigated presently exhibit intrasession and intersession variability approximately less than 5% CoV, interscanner effects of approximately 5%-10% CoV, and intersubject effects of approximately 5%-15% CoV. 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



**FIGURE 7** Variability in connectomics. Visualization of variation across intrasession, intersession, interscanner, and intersubject groups illustrates increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon rank-sum test with and without Bonferroni correction.



**FIGURE 8** Overall trends in coefficient of variation (CoV) across DTI, NODDI, bundle segmentation, and connectomics. Visualization of median CoV across all four processing approaches on the intrasession, intersession, interscanner, and intersubject levels illustrates consistently increased variability with session, scanner, and subject effects. Statistical significance was determined with the Wilcoxon signed-rank test with and without Bonferroni correction. The outlying points correspond to the NODDI cVF approach in white matter where absolute cVF values are expected to be low.

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,<sup>50–53</sup> multicompartiment models,<sup>53,54</sup> tractography and bundle segmentation,<sup>55,56</sup> and connectomics<sup>57,58</sup> (Table 1). Looking at the literature, we find many existing studies used CoV to estimate variability. Thus, we elected to center our study around this approach to better place our results in the context of the literature. We found similar estimates of variability between our results and those of prior studies. However, review of the literature also demonstrates a fragmented picture of DWI variability. Previous studies have largely each focused on one type of approach and one or two levels of variation. This, coupled with different definitions of variability and different study objectives, has 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 to characterize all four types of diffusion processing and all four levels of variation consistently and simultaneously. Thus, we hope that the data set 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.

For this study, we chose popular software toolboxes to do all of 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,<sup>59–61</sup> for identifying regions<sup>48,62–64</sup> and bundles,<sup>65–68</sup> for comparing bundles,<sup>69</sup> and for configuring and representing connectomes.<sup>31,58,70,71</sup> Additionally, there are a number of other microstructural measures that can be characterized as well.<sup>19</sup> Thus, the goal of the present study was not to provide an analysis between different processing toolboxes or parameters, and because each approach was not necessarily optimized, we do not recommend thorough use of the 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 data set. 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 data sets and studies for variability use adult subjects. Of the existing pediatric data sets, many have focused on older age ranges. For example, the Adolescent Brain Cognitive Development project<sup>72</sup> and the Lifespan Human Connectome Project in Development<sup>73</sup> 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 data sets of children before 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 an analogous characterization of the longitudinal intersession variability in children with 1 year between sessions compared with the adult intersession variability computed previously for all four processing approaches (Supporting Information Figure S2). We hope that investigators in developmental neuroscience and pediatric neurology will be able to take advantage of this resource for their work.

We note that the groups in each of the variability levels described in this study are necessarily distributed across different nested effects. For instance, because sessions are nested in

scanners which are nested in subjects, the intrasession groups are distributed across different sessions, scanners, and subjects; the intersession groups are distributed across different scanners and subjects, and so forth. Thus, one limitation of our study is that in an effort to better place our results in context of the literature with interpretable metrics like CoV, we partially but not fully isolate the appropriate session, scanner, and subject biases. Similarly, another limitation of our study is the differences in the number of gradient directions between the different cohorts. Cohort III consists of a 40-direction  $b = 1000 \text{ s/mm}^2$  acquisition and a 56-direction  $b = 2000 \text{ s/mm}^2$  acquisition in contrast to the 96 directions for cohorts I and II. This is a potential effect that could be biasing the results. Similarly, due to hardware limitations, the data collected at site 3 in cohort II were collected at a maximum shell of  $2465 \text{ s/mm}^2$ , as opposed to the  $2500 \text{ s/mm}^2$  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 data set. Thus, considering these potential effects, future directions include developing a mixed-effects model capable of estimating variability in an interpretable manner as well as robustly modeling the nested nature of sessions, scanners, and subjects and the acquisition biases.

## ACKNOWLEDGMENTS

The authors thank E. Brian Welch for his help with image acquisition and study design, Zachary J. Williams for his statistical insight, 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, Tennessee.

## DATA AVAILABILITY STATEMENT

We have made the MASiVar data set publicly available at <https://openneuro.org/datasets/ds003416> in Brain Imaging Data Structure (BIDS) format with de-identified metadata and defaced images.<sup>74</sup>

## ORCID

Leon Y. Cai  <https://orcid.org/0000-0002-5812-5397>

Benjamin N. Conrad  <https://orcid.org/0000-0003-3755-6988>

Bennett A. Landman  <https://orcid.org/0000-0001-5733-2127>

## TWITTER

Leon Y. Cai  @LeonCai

## REFERENCES

- O'Donnell LJ, Westin CF. An introduction to diffusion tensor image analysis. *Neurosurg Clin N Am*. 2011;22:185-196.
- Assaf Y, Pasternak O. Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review. *J Mol Neurosci*. 2008;34:51-61.
- Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. *Biophys J*. 1994;66:259-267.
- 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.
- 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.
- 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.
- 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.
- Sotiropoulos SN, Zalesky A. Building connectomes using diffusion MRI: why, how and but. *NMR Biomed*. 2019;32:1-23.
- 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.
- Travers BG, Adluru N, Ennis C, et al. Diffusion tensor imaging in autism spectrum disorder: a review. *Autism Res*. 2012;5:289-313.
- Repple J, Karliczek G, Meinert S, et al. Variation of HbA1c affects cognition and white matter microstructure in healthy, young adults. *Mol Psychiatry*. 2021;26:1399-1408.
- 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.
- De Santis S, Bastiani M, Droyb 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.
- 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*. 2020;25:3208-3219.
- 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.
- 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.
- 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.
- 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.
- Koller K, Rudrapatna U, Chamberland M, et al. MICRA: microstructural image compilation with repeated acquisitions. *Neuroimage*. 2020;225:117406.

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.
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;7:157.
22. 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.
23. Cai LY, Yang QI, Hansen CB, et al. PreQual: an automated pipeline for integrated preprocessing and quality assurance of diffusion weighted MRI images. *Magn Reson Med*. 2021;86:456-470.
24. 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.
25. Veraart J, Fieremans E, Novikov DS. Diffusion MRI noise mapping using random matrix theory. *Magn Reson Med*. 2016;76:1582-1593.
26. Cordero-Grande L, Christiaens D, Hutter J, Price AN, Hajnal JV. Complex diffusion-weighted image estimation via matrix recovery under general noise models. *Neuroimage*. 2019;200:391-404.
27. 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.
28. 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.
29. 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.
30. Garyfallidis E, Côté M-A, Rheault F, et al. Recognition of white matter bundles using local and global streamline-based registration and clustering. *Neuroimage*. 2018;170:283-295.
31. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*. 2010;52:1059-1069.
32. Hollander M, Wolfe DA, Chicken E. *Nonparametric Statistical Methods*. Hoboken, New Jersey: John Wiley & Sons; 2013.
33. 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.
34. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297-302.
35. 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.
36. Westin CF, Peled S, Gudbjartsson H, Kikinis R, Jolesz FA. Geometrical diffusion measures for MRI from tensor basis analysis. In: Proceedings of the 5th Annual Meeting of ISMRM, Vancouver, Canada, 1997. Abstract #1742.
37. Tustison NJ, Cook PA, Klein A, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*. 2014;99:166-179.
38. 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.
39. Mori S, Wakana S, Van Zijl PCM, Nagae-Poetscher LM. *MRI Atlas of Human White Matter*. Amsterdam, Netherlands: Elsevier; 2005.
40. Wakana S, Caprihan A, Panzenboeck MM, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage*. 2007;36:630-644.
41. 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.
42. Garyfallidis E, Brett M, Amirbekian B, et al. Dipy, a library for the analysis of diffusion MRI data. *Front Neuroinform*. 2014;8:1-17.
43. 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.
44. Tournier J-D, Smith R, Raffelt D, et al. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage*. 2019;202:116137.
45. Smith RE, Tournier JD, Calamante F, Connelly ASIFT. Spherical-deconvolution informed filtering of tractograms. *Neuroimage*. 2013;67:298-312.
46. 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.
47. 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.
48. 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.
49. Goldstein JM, Seidman LJ, Makris N, et al. Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability. *Biol Psychiatry*. 2007;61:935-945.
50. Magnotta VA, Matsui JT, Liu D, et al. Multicenter reliability of diffusion tensor imaging. *Brain Connect*. 2012;2:345-355.
51. 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.
52. Palacios EM, Martin AJ, Boss MA, et al. Towards precision and reproducibility of DTI. *AJNR Am J Neuroradiol*. 2017;38:537-545.
53. 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.
54. Tariq M, Schneider T, Alexander DC, Wheeler-Kingshott CA, Zhang H. Assessing scan-rescan reproducibility of the parameter estimates from NODDI. In: Proceedings of the 21st Annual Meeting of ISMRM, Salt Lake City, Utah, 2013. Abstract #3187.
55. Besseling RMH, Jansen JFA, Overvliet GM, et al. Tract specific reproducibility of tractography based morphology and diffusion metrics. *PLoS One*. 2012;7:e34125.
56. 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.
57. Prčkovska V, Rodrigues P, Puigdellivol Sanchez A, et al. Reproducibility of the structural connectome reconstruction across diffusion methods. *J Neuroimaging*. 2016;26:46-57.
58. Roine T, Jeurissen B, Perrone D, et al. Reproducibility and inter-correlation of graph theoretical measures in structural brain connectivity networks. *Med Image Anal*. 2019;52:56-67.
59. 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.
60. Chang LC, Jones DK, Pierpaoli C. RESTORE: robust estimation of tensors by outlier rejection. *Magn Reson Med*. 2005;53:1088-1095.
  61. Cook PA, Bai Y, Seunarine KK, Hall MG, Parker GJ, Alexander DC. Camino: open-source diffusion-MRI reconstruction and processing. In: Proceedings of the 14th Annual Meeting of ISMRM, Seattle, Washington, 2006. Abstract #2759.
  62. 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.
  63. 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.
  64. 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.
  65. 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.
  66. Yeh FC. Shape analysis of the human association pathways. *Neuroimage*. 2020;223:117329.
  67. 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.
  68. 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.
  69. 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.
  70. Hagmann P, Cammoun L, Gigandet X, et al. Mapping the structural core of human cerebral cortex. *PLoS Biol*. 2008;6:1479-1493.
  71. Sporns O, Tononi G, Kötter R. The human connectome: a structural description of the human brain. *PLoS Comput Biol*. 2005;1:245-251.
  72. 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.
  73. 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.
  74. 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.
  75. Grech-Sollars M, Hales PW, Miyazaki K, et al. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. *NMR Biomed*. 2015;28:468-485.
  76. Cousineau M, Jodoin P-M, Garyfallidis E, et al. A test-retest study on Parkinson's PPMI dataset yields statistically significant white matter fascicles. *NeuroImage Clin*. 2017;16:222-233.
  77. 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*. 2018;12:1055.
  78. 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.
  79. Girard G, Whittingstall K, Deriche R, Descoteaux M. Studying white matter tractography reproducibility through connectivity matrices. In: Proceedings of the 23rd Annual Meeting of ISMRM, Toronto, Canada, 2015.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

**FIGURE S1** Volume-to-volume motion and SNR of raw data in MASiVar cohorts by site

**FIGURE S2** Comparison of adult intersession variability to pediatric intersession variability with approximately 1 year between sessions in children aged 5-8 years. Statistical significance was determined with the Wilcoxon rank-sum test

**FIGURE S3** Example identification of scan pairs at the four levels of variation. The MASiVar data set 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 S4** Defining variability within a pair of images. A, Regional fractional anisotropy (FA), mean diffusivity (MD), CSF volume fraction (cVF), intracellular volume fraction (iVF), and orientation dispersion index (ODI) variability was defined as the median voxel-wise absolute percent difference. Absolute angular difference was used for principal eigenvector (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 maximum modularity (MM), global efficiency (GE), and characteristic path length (CPL) graph measures was defined with absolute percent difference (2-4)

**FIGURE S5** Summarizing paired variability at a given level of variation. A, The DTI, neurite orientation dispersion and density imaging (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 S6** Paired variability in DTI, NODDI, bundle segmentation, and connectomics. Visualization of DTI (A)

and NODDI (B) 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. C, Visualization of bundle segmentation 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. D, With the exception of the intersession and interscanner correlation comparison, visualization of connectomics differences within intrasession, intersession, interscanner, and intersubject pairs consistently illustrates increased variability with session, scanner, and subject effects. Statistical

significance was determined at 0.05 (0.008 Bonferroni-corrected) with six pair-wise Wilcoxon signed-rank tests for the DTI, NODDI, and bundle comparisons and with Wilcoxon rank-sum tests for connectomics. The *P*-values reported are uncorrected

**TABLE S1** List of 43 white-matter bundles investigated with the RecoBundles method

**How to cite this article:** Cai LY, Yang Q, Kanakaraj P, et al. MASiVar: Multisite, multiscanner, and multisubject acquisitions for studying variability in diffusion weighted MRI. *Magn Reson Med*. 2021;86:3304–3320. <https://doi.org/10.1002/mrm.28926>