
Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Fiber tractography bundle segmentation depends on scanner effects, vendor effects, acquisition resolution, diffusion sampling scheme, diffusion sensitization, and bundle segmentation workflow

Kurt G. Schilling*a,b, Chantal M.W. Taxc, Francois Rheaультd, Colin Hansend, Qi Yangd, Fang-Cheng Yehc, Leon Caид, Adam W. Andersona,b,cf, Bennett A. Landmana,b,df

*a Department of Radiology & Radiological Science, Vanderbilt University Medical Center, Nashville, TN, United States
b Department of Neurological Surgery, University of Pittsburgh, United States
c Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff University, Cardiff, United Kingdom
d Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, United States
e Department of Neurology, University of Texas Southwestern Medical Center, Dallas, TX, United States
f Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States

ARTICLE INFO

Keywords:
Tractography
Bundle segmentation
White matter
Reproducibility
Harmonization

ABSTRACT

When investigating connectivity and microstructure of white matter pathways of the brain using diffusion tractography bundle segmentation, it is important to understand potential confounds and sources of variation in the process. While cross-scanner and cross-protocol effects on diffusion microstructure measures are well described (in particular fractional anisotropy and mean diffusivity), it is unknown how potential sources of variation effect bundle segmentation results, which features of the bundle are most affected, where variability occurs, nor how these sources of variation depend upon the method used to reconstruct and segment bundles. In this study, we investigate six potential sources of variation, or confounds, for bundle segmentation: variation (1) across scan repeats, (2) across scanners, (3) across vendors (4) across acquisition resolution, (5) across diffusion schemes, and (6) across diffusion sensitization. We employ four different bundle segmentation workflows on two benchmark multi-subject cross-scanner and cross-protocol databases, and investigate reproducibility and biases in volume overlap, shape geometry features of fiber pathways, and microstructure features within the pathways. We find that the effects of acquisition protocol, in particular acquisition resolution, result in the lowest reproducibility of tractography and largest variation of features, followed by vendor-effects, scanner-effects, and finally diffusion scheme and b-value effects which had similar reproducibility as scan-rescan variation. However, confounds varied both across pathways and across segmentation workflows, with some bundle segmentation workflows more (or less) robust to sources of variation. Despite variability, bundle dissection is consistently able to recover the same location of pathways in the deep white matter, with variation at the gray matter/white matter interface.

Next, we show that differences due to the choice of bundle segmentation workflows are larger than any other studied confound, with low-to-moderate overlap of the same intended pathway when segmented using different methods. Finally, quantifying microstructure features within a pathway, we show that tractography adds variability over-and-above that which exists due to noise, scanner effects, and acquisition effects. Overall, these confounds need to be considered when harmonizing diffusion datasets, interpreting or combining data across sites, and when attempting to understand the successes and limitations of different methodologies in the design and development of new tractography or bundle segmentation methods.

1. Introduction

Diffusion-weighted magnetic resonance imaging (dMRI) has proven valuable to characterize tissue microstructure in health and disease (Alexander et al., 2019; Jones et al., 2018; Novikov et al., 2018). Moreover, the use of dMRI fiber tractography to virtually dissect fiber pathways (Jeurissen et al., 2019) is increasingly used to localize microstructure measurements to specific white matter bundles (Raffelt et al., 2017; Chamberland et al., 2019; Yeatman et al., 2012), and to study the connections and shapes of pathways (Jeurissen et al., 2019; Maffei et al., 2019; Forkel et al., 2014; Hau et al., 2017; Hau et al., 2016; Sarubbo et al., 2019; Sarubbo et al., 2013; Neubert et al., 2015; Neubert et al., 2014; Mars et al., 2012). Despite promises of noninvasive measurements of white matter features, variability may exist in measurements due to inherent variability within scanners and across.
scanners, differences in acquisition protocol parameters, and differences due to processing pipelines, amongst others. These sources of variance challenge the quantitative nature of derived measures of microstructure and connectivity, and hinder the ability to interpret different findings or combine different datasets.

These effects have been intensively studied for tissue microstructure features, specifically diffusion tensor imaging (DTI) (Pierpaoli et al., 1996) indices of fractional anisotropy (FA) and mean diffusivity (MD). Numerous studies have characterized intra-scanner and inter-scanner DTI variability (AK Prohl et al., 2019; Mirzaalian et al., 2016; VA Magnotta et al., 2012; Landman et al., 2011; Teipel et al., 2011; Vollmar et al., 2010; Farrell et al., 2007; Landman et al., 2007; Heiervang et al., 2006; Pfefferbaum et al., 2003; Jones, 2003; Lori et al., 2002), differences due to acquisition parameters (Farrell et al., 2007; Landman et al., 2007; Jones, 2003; Jones et al., 2020; Papinutto et al., 2013; Jones, 2004; Jones and Basser, 2004) including image resolution, number of diffusion images, and diffusion sensitization (i.e. the b-value), and differences due to processing and algorithmic choices (Jones et al., 2007; Chang et al., 2005). These have paved the way towards recommendations and guidelines for reliable and reproducible DTI (Jones et al., 2013; Jones, 2010; Jones and Cercignani, 2010; DK Jones et al., 1999); however, a standardized universal dMRI protocol does not exist, and differences are expected across sites and studies (Fig. 1) (L. Ning et al., 2020; CM Tax et al., 2019). Yet, there is significant interest in combining data from different sites to increase statistical power and benefit from multi-center recruitment abilities (Mirzaalian et al., 2016; L. Ning et al., 2020; Zhong et al., 2020; Cetin Karayumak et al., 2019; KM Huynh et al., 2019; Vishwesh Nath et al., 2018; Yu et al., 2018; Mirzaalian et al., 2018; Fortin et al., 2017), and it is clear that these differences need to be accounted for, or removed, prior to data aggregation or joint statistical analysis.

Notwithstanding the increased awareness and improved characterization of dMRI microstructural measures, very little work has been performed to characterize and understand reproducibility of tractography-derived features across scanners, across protocols, and across different tractography bundle segmentation algorithms (Pestilli et al., 2014; Nath et al., 2019). Variability in tractography estimates of fiber pathways will further increase variability in connectivity analyses and impact microstructural characterization, e.g. when tractography is used to define ROIs or to perform along-tract profiling. While few studies do exist, they are often limited to a single pathway (Chamberland et al., 2018; F Rheault et al., 2020), a single dissection protocol (Vazquez et al., 2020; Guevara et al., 2017), or a single source of potential variation (Schilling et al., 2020), such as test-retest or population-based reproducibility (Guevara et al., 2017; F Zhang et al., 2019; Guevara et al., 2012). Additionally, they do not investigate where in the brain or along...
the pathway that this variability occurs, and are often limited to charac-
terizing only microstructural features of these pathways (i.e., the FA or MD along or within the pathway) (Heier et al., 2006; Wakana et al., 2007). Thus, we currently do not which sources of variation impact tractography bundle segmentation the most, which features of the bundle are most affected, where variability occurs, nor how these questions are dependent upon the workflow used to dissect fiber bundles. Thus, for the first time, we combine, assess, and rank all previously studied sources of potential variation in the same study, with a focus on tractography rather than just DTI measures.

Here, we investigate and compare the reproducibility of tractography across six confounds, or sources of variation: intrinsic variability across scan repeats, differences across scanners, across vendors, across different acquisition spatial resolution and acquisition angular resolution, and across different diffusion sensitizations (b-values). We employ and examine four fully-automated and commonly utilized bundle reconstruction workflows on two cross-scanner cross-protocol benchmark datasets. We first investigate how these confounds affect not only the overlap and location of pathways, but also evaluate variability in topological measures of the bundle including length, area, shape, and volume features. We ask which pathways, which bundle segmentation workflow, and which features are most reproducible? And what source of variation is most significant for each method? Second, we visualize where in the brain, and where within a pathway, tractography is most vari-
able (and most robust) and investigate if sources of variation effect this in different ways. Third, we quantify and visualize differences in tractography that result when using different bundle segmentation workflows. Finally, we analyze traditional DTI measures and quantify differences due to these sources of variation as well as the added variance intro-
duced by the tractography process over and above that inherent across scanners and across acquisition protocols.

2. Methods

2.1. Datasets

Here we utilize two open-sourced multi-subject, multi-scanner, and multi-protocol benchmark databases: the MAStiVr (Cai et al., 2020) and MUSHAC datasets (L. Ning et al., 2020; CM Tax et al., 2019). We note that other multi-site databases exist (see Discussion), although they are often limited to investigating differences across subjects and scanners, whereas the two chosen datasets together allow investigation of repeats, scanners, vendors, and acquisition protocols (resolution, direction, b-values).

2.1.1. MUSHAC dataset

The MUSHAC database will allow investigation of cross-scanner, cross-protocol, and cross b-value effects (L. Ning et al., 2020; CM Tax et al., 2019). This database was part of the 2018 and 2019 MICCAI Harmonization challenge. Here, we utilize the data acquired from 10 healthy subjects used as training data in the challenge, and described in (L. Ning et al., 2020; CM Tax et al., 2019). Each subject has 4 unique datasets. This work focuses on the data acquired on two scanners with different gradient strengths: a) 3T Siemens Prisma (80 mT/m), and b) 3T Siemens Connectom (300 mT/m). Two types of protocols were acquired from each scanner: 1) a ‘standard’ protocol with acquisition parameters matched to a typical clinical protocol; and 2) a more advanced or ‘state-of-the-art’ protocol where the superior hardware and software specifications were utilized to increase the number of acquisitions and spatial resolution per unit time. The ‘standard’ protocol from both scanners is matched as closely as possible, with an isotropic resolution of 2.4 mm, TE=89 ms and TR=7.2 s, and 30 diffusion-weighted directions acquired at two b-values: b = 1200, 3000 s/mm2 (scan time ~7.5 min). On the other hand, the Prisma ‘state-of-the-art’ data has a higher isotropic res-
olution of 1.5 mm, TE=80 ms, TR=7.1 s and 60 directions at the same b-values (~14.5 min). While the Connectom ‘state-of-the-art’ data has the highest resolution of 1.2 mm with TE=68 ms, TR=5.4 s and 60 di-
rections (~11 min). All data was corrected for distortions, motion, eddy currents (Andersson et al., 2003), and gradient nonlinearity distortions (Gasser et al., 2013). For each subject, the Prisma standard-acquisition dataset was used as a reference space and all additional datasets were affinely registered to this space using the corresponding FA maps with FSL Flirt with appropriate b-vector rotation.

2.1.2. MAStiVr dataset

The MAStiVr database will allow investigation of scan-rescan and cross-scanner effects. Here we used a subset of Cohort II of this database described in (Cai et al., 2020), which consisted of 5 healthy subjects with 6 unique “datasets”. Each subject was scanned on four scanners: a) 3T Philips Achieve (80 mT/m) and b) a different 3T Philips Achieve (60mT/m) at the same site, c) a 3T General Electric Discov-
ery MR750 Scanner at a different site, and d) a 3T Siemens Skyra scanner at a different site. These acquisitions were matched as closely as possible and are similar to that of the standard-protocol described above: with an isotropic resolution of 2.5 mm, TE=55 ms and TR=6.2 s (7.0 s for scanner-b), and 32 diffusion-weighted directions acquired at b = 1000s/mm2 (scan time ~3.5 min). Additionally, the subjects were scanned twice on the first scanner, and also had an acquisition that con-
sisted of a 96-direction b = 1000 dataset, both of which were also utilized in the current study. We note that one subject did not have a repeat scan on the first scanner (a) and one subject did not have a scan on the GE Scanner (b).

All data were corrected for distortions, motion, and eddy currents (Andersson et al., 2003; Cai et al., 2021). For each subject, the first session on scanner-a was used as a reference space and all additional datasets were affinely registered to this space using the corresponding FA maps with FSL Flirt (Jenkinson et al., 2012) with appropriate b-vector rotation.

2.2. Sources of variation

We investigate several possible sources of variation in the bundle segmentation process.

RESSCAN: the effects of repeating a scan on the same scanner (i.e. scan-rescan) in a different session, but with a matched acquisition. This effect is quantified using the repeated acquisitions from the MAStiVr database.

SCAN1: inter-scanner (cross-scanner) effects, with a matched acqui-
sition and of the same vendor. SCAN1 is quantified using the matched acquisitions from the MAStiVr database acquired on different Philips scanners (both Philips Achieva).

SCAN2: inter-scanner (cross-scanner) effects, with a matched acqui-
sition and of the same vendor. SCAN2 is quantified using the matched standard acquisitions from the MUSHAC acquired on different Siemens scanners (Siemens Connectom and Siemens Prisma).

VEN1: inter-vendor (cross-vendor) effects, with a matched acqui-
sition. VEN1 is quantified using the matched acquisitions from the MAStiVr database, but acquired on scanners from different vendors (Philips Achieva and General Electric Discovery).

VEN2: inter-vendor (cross-vendor) effects, with a matched acqui-
sition. VEN2 is quantified using the matched acquisitions from the MAStiVr database, but acquired on scanners from different vendors (Philips Achieva and Siemens Skyra).

RES1: effects of spatial resolution, with matched scanner, diffusion directions, and b-value. RES1 is quantified by using the MUSHAC acquisi-
tions from the Prisma standard-acquisition and from the Prisma state-
of-the-art acquisition but with only 30 uniformly distributed directions utilized (to match the standard-acquisition). This represents differences between a 2.4 mm isotropic and 1.5 mm isotropic acquisition.

RES2: effects of spatial resolution, with matched scanner, diffusion directions, and b-value. RES2 is quantified by using the MUSHAC acquisi-
tions from the Connectom standard-acquisition and from the Conne-
tom state-of-the-art acquisition but with only 30 uniformly distributed directions utilized (to match the standard-acquisition). This represents differences between a 2.4 mm isotropic and 1.2 mm isotropic acquisition.

DIR1: effects of number of diffusion-weighted directions, with matched scanner, resolution, and b-value. DIR1 is quantified using the MASHvar acquisitions from the first scanner at 32 directions and the acquisition on the same scanner at 96 directions.

DIR2: effects of number of diffusion-weighted directions, with matched scanner, resolution, and b-value. DIR2 is quantified using the MUSHAC acquisitions from the state-of-the art Prisma acquisition with only 30 uniformly distributed directions utilized and the full state-of-the art acquisition which consists of 60 directions.

BVAL: effects of changing the b-value, on the MUSHAC Prisma scanner with the ‘standard’ protocol, from $b = 1200$ to $b = 3000$, within the same acquisition.

We note that we also investigated a second effect of b-value (within the state-of-the art Prisma protocol, with no statistically significant differences, and for figure simplicity only show the above-mentioned b-value analysis). Previous version of this manuscript (and preprint) included an ACQ1 and ACQ2 (from state-of-the art to standard-acquisition) that were isolated into both effects of directions (DIR1 and DIR2) and resolution (RES1 and RES2).

A final source of variation investigated is that caused by the use of different bundle reconstruction workflows. Because all workflows segment different numbers of, and sets of, fiber pathways (see below), for this analysis, we investigated only those fiber pathways which are common to all algorithms. In this case, we identified 7 (bilateral) pathways which are segmented by all automated methods.

### 2.3. Tractography bundle dissection

We utilized four common, well-validated, and fully-automated fiber bundle reconstruction workflows, all implemented using standard and/or recommended settings. It is important to highlight that each workflow included differences in local fiber-direction estimation, fiber tractography, and bundle segmentation algorithms, and our attempt was to implement the entire workflow as would be done in a typical scientific study (see Discussion on limitations of confounds due to differences in bundle segmentation process). While there are dozens of bundle segmentation algorithms, we have chosen these to be representative of common approaches, utilizing regions of interest, atlases, machine learning, templates, etc. (see Discussion and Limitations).

#### 2.3.1. TractSeg

TractSeg is based on convolutional neural networks and performs bundle-specific tractography based on a field of estimated fiber orientations (Wasserthal et al., 2019; J Wasserthal et al., 2018; J Wasserthal et al., 2018). We implemented the dockerized version at (https://github.com/MIC-DKFZ/TractSeg), which generates fiber orientations using constrained spherical deconvolution with the MRtrix3 software (Tournier et al., 2019). We note that different reconstruction methods could have been chosen to generate fiber orientations. This method dissects 72 bundles.

#### 2.3.2. Automatic fiber tractography (ATK)

ATK was performed in DSI Studio software using batch automated fiber tracking (Yeh, 2020). Data were reconstructed using generalized q-sampling imaging (Yeh et al., 2010) with a diffusion sampling

---

*Fig. 2. Tractography varies across scanners, acquisitions, b-values, and bundle segmentation methods. On the same subject, the arcuate fasciculus is shown for each of the 4 bundle segmentation methods, for two scanners and two acquisitions. Note that the pathway is visualized as streamlines for TractSeg, ATK, and Reco but a probability density map for Xtract. Arrows highlight visible examples of differences in streamlines across scanners (solid arrows), across acquisition (dotted arrows), and across b-values (dashed arrows).*
length ratio of 1.25. 20 white matter pathways were automatically reconstructed using seeding regions defined in the HCP842 tractography atlas (Yeh et al., 2018), randomly generated tracking parameters of anisotropy threshold, angular threshold, step size, and subsequent segmentation and pruning. The Dockerized source code is available at http://dsi-studio.labsolver.org.

2.3.3. Recobundles (RECO)

Recobundles (Garyfallidis et al., 2018) segments streamlines based on their shape-similarity to a dictionary of expertly delineated model bundles (Yeh et al., 2018). Recobundles was run using DIPY (Garyfallidis et al., 2014) software (https://dipy.org) after performing whole-brain tractography using spherical deconvolution and DIPY LocalTracking algorithm. The bundle-dictionary contains 80 bundles, but only 44 were selected to be included in this study after consulting with the algorithm developers based on internal quality assurance (for example, removing cranial nerves which are often not used in brain imaging). Of note, Recobundles is a method to automatically extract and recognize bundles of streamlines using prior bundle models, and the implementation we chose uses the DIPY bundle dictionary (Yeh et al., 2018) for extraction, although others can be used, as well as alternative shape-similarity filtering criteria.

2.3.4. Xtract

Xtract (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/XTRACT) is a recent automated method for probabilistic tractography based on carefully selected inclusion, exclusion, and seed regions, defined in a standard space (Warrington et al., 2020). Xtract used the ball-and-stick model of diffusion from FSL’s bedpostx algorithm (Jenkinson et al., 2012), in combination with a probabilistic tractography algorithm probtrackx, to reconstruct 42 white matter pathways. In contrast to the preceding methods, which result in streamlines, this method results in visitation count maps for each pathway.

A list of all segmentations generated from each method and corresponding acronyms is given in the appendix. The 7 pathways identified to be common to all tractography bundle segmentation techniques includes: arcuate fasciculus (AF), corticospinal tract (CST), inferior fronto-occipital fasciculus (IFO), inferior longitudinal fasciculus (ILF), middle longitudinal fasciculus (MidLF), optic radiations (OR), and uncinate fasciculus (UF), all of which are bilateral including left (_L) and right (_R) hemisphere pathways.

A thorough quality control was performed for all subjects, and for all pathways. This included first visualization and verification of adequate alignment of all FA maps (to ensure appropriate quantification of overlap measures). Second, all pathways, for a subjects,
were visualized in mosaic form using tools from the SCILPY toolbox (https://github.com/scilus/scilpy), and pathways were visually assessed and removed from analysis if deemed in the incorrect location or shape. Finally, individual bundles were removed from analysis if the number of segmented streamlines was less than 3 standard deviations away from the mean number (for each pathway), or if the total number of streamlines was below 200 (indicating failure of tractography), and subjects were removed from analysis (for a given algorithm) if >20% of pathways failed QC.

2.4. Feature extraction

A number of features were extracted from each bundle segmented. First, for simple comparisons of the volume occupied by each pathway, all bundles (from all methods) were binarized and resampled at 1 mm isotropic resolution. For methods generating streamlines (Tractseg, ATK, and RECO) this is equivalent to binarizing based on a streamline density of 1. Because Xtract output is in the form of a normalized probability distribution, where a threshold of 2.5E-4 was chosen based on (Warrington et al., 2020). The binarized segmentation was used for measures of Dice overlap (described below).

Second, several descriptors of the shape and geometry of the bundles were extracted. Shape analysis was performed using DSI Studio, and made available as matlab code (https://github.com/dmitrishastin/tractography_shapes/), based on (Yeh, 2020), to derive length, area, volume, and shape metrics of a bundle. Briefly, length features include mean length, span, diameter, and average radius of end regions. Area features include total surface area and the total area of end regions. Volume features include total volume, trunk volume, and branch volume. Shape features include pathway curl, elongation, and irregularity.

Finally, microstructure measures of FA and MD (calculated using iteratively reweighted linear least squares estimator) within pathways were extracted. In all cases, a simple measure of the average value within the binary volume was performed, although we note that these measures can also be weighted by certainty or streamline density.
isolate the added variation due to tractography from that of the existing sources of variation, these measures were extracted in two ways. First, using the binary regions defined in the reference scan-space only (i.e., the Prisma standard-acquisition and first session on scanner-a for MUSHAC and MASIvar datasets respectively) were used as the same region-of-interest across all effects, in order to isolate each source of variation while keeping ROIs constant. Second, the binary region defined by tractography for each specific dataset was used to extract the average FA (or MD), which includes both variation due to the effect under investigation and the variation due to tractography differences.

2.5. Reproducibility evaluation

Reproducibility was evaluated using several metrics, and across each source of variation. First, the Dice overlap was calculated for each pair of bundles as an overall measure of similarity of volumes. The Dice overlap is calculated as twice the intersection divided by the sum of the volumes of each dataset. Results were displayed across all fiber pathways for a given source of variation, and differences between effects were calculated using the nonparametric paired (i.e. same subject, different effect) Wilcoxon signed rank tests.

Differences in scalar shape features are calculated as the mean absolute percentage error (MAPE), sometimes referred to as the mean absolute percentage deviation. For two different scans, this measure is calculated as the difference divided by the mean, and can be converted to a percentage error by multiplication by 100. This measure was calculated over all subjects, and results were displayed across all fiber pathways for a given source of variation. Differences between effects were again calculated using the nonparametric paired (i.e. same subject, different effect) Wilcoxon signed rank tests.

For visual comparisons only, all subjects were nonlinearly registered to MNI space, using the 1 mm isotropic FA template and the corresponding FA maps with FSL FLIRT + FNIRT. Streamlines were directly warped to this space for visualization of agreement/disagreement across the cohort. Note that quantification of shape features was performed in native space prior to warping.

For all statistical analysis, thresholds were corrected for multiple comparisons. For example, when investigating differences in effects of DICE/MAPE, etc., we tested differences between 10 effects, resulting in 55 tests performed for each analysis.

3. Results

3.1. Qualitative variation

Fig. 1 shows FA maps of the same subject, but acquired on different scanners and with different protocols. In agreement with the literature (L. Ning et al., 2020; CM Tax et al., 2019), differences in magnitude, contrast, and signal-to-noise ratios are readily apparent, and dMRI measures qualitatively vary due to scanner and acquisition effects.

Fig. 2 shows tractography bundle segmentation results for an example pathway (the arcuate fasciculus; AF) on a single subject, for two scanners, two protocols, two b-values, and all four reconstruction methods. For a given bundle segmentation method, minor differences are observed in individual gyri and at regions of low streamline density. However, bundles are visually very similar across scanners and protocols, with similar shapes, locations, curvatures, and connections. Most notably, and as expected (Schilling et al., 2020), the biggest differences are observed when comparing the same pathway across different bundle segmentation methods.

3.2. Quantitative variation due to rescan, scanner, vendor, resolution, directions, and b-value effects

The effects of RESCAN, SCAN1, SCAN2, VEN1, VEN2, DIR1, DIR2, and BVAL on Dice overlap coefficient is shown in Fig. 3 for fourteen selected pathways common to all bundle segmentation methods. Notably, reproducibility is most dependent on the bundle dissection method, with TractSeg consistently resulting in high reproducibility for all sources of variation. Within a method, most pathways show similar patterns of reproducibility. For example, for TractSeg and Xtract all pathways indicate high RESCAN, DIR(1 and 2) and BVAL reproducibility, but are most sensitive to RES, with RES2 showing more variation than RES1. Additionally, Dice overlap shows some variation across pathways, for example CST and UF generally have higher overlap than OR, IFO, and AF, although trends are different for different workflows.

The results of the Dice overlap coefficient-analysis for each method is shown in Fig. 4, but condensed across all pathways within a given bundle segmentation method. Similar trends are observed as in Fig. 3, with TractSeg consistently indicating the highest Dice overlap, and all methods indicating moderate-to-good overall overlap for most pathways.

![Fig. 5. Locations of agreement and disagreement across effects. Maps are computed by overlaying (for each source of variation), maps of where there is overlap (i.e. agreement) and non-overlap (disagreement), averaged across all subjects. For each effect, the percent agreement indicates areas where a pathway is consistently located and is shown using a “hot” colormap, while the percent disagreement indicates areas without consistent overlap and is shown using a “cold” colormap. Results are shown for a highly reproducible pathway (AF, disected using TractSeg) and for a less reproducible pathway (SLF2 dissected using XTRACT). Note that even though disagreement is abundant, it does not consistently occur (i.e., % disagreement remains low; black and dark blue) suggesting no systematic bias due to effects, and disagreements are largely attributed to the stochastic nature of the tractography and dissection process.](image-url)
For the less reproducible pathway, the agreement is moderate to high in the dense core, or center, of the pathway in the deep white matter. Again, disagreements are at the edges, and prominent at the white matter and gray matter boundary. However, even though disagreement is more noticeable, the percent-disagreement remains low, indicating random disagreement as opposed to a consistent bias in the spatial location of this pathway. In this case, sources of variation from SCAN2 and RES2 and VEN1 are more noticeable as a larger source of variation, in agreement with quantitative results.

3.4. Variation of shape features

Fig. 7 summarizes the MAPE of different features across different sources of variation. Again, Curl, Length, and Span are highly reproducible across all effects, with MAPE always below 10%, and surface area and volume result in higher MAPE. Trends are the same as those observed for Dice overlap, with generally larger differences due to resolution and vendor acquisition effects (RES 1 and 2, VEN 1 and 2), followed by scanner effects (SCAN1 showing the largest variation).

3.3. Localization of variation

Fig. 5 visualizes locations of tractography bundle segmentation agreement (or consistency), and where it disagrees (variability) as hot and cold color maps, respectively. Agreement and disagreement are averaged across all subjects and shown for all sources of variation. For display, we have chosen an example pathway that is highly reproducible (the AF from TractSeg) and one which displayed lower reproducibility (the SLFII from Xtract). For the highly reproducible pathway, all sources of variation show very similar results. The agreement is very high throughout the entire pathway (hot colors), and percent-disagreement remains fairly low (black and dark blue colors). This means that when two bundles disagree, the disagreement is largely randomly distributed, rather than a consistent localized bias introduced by a certain source of variation—an effect which would show up as a consistent disagreement (i.e., a high percent-disagreement). Disagreement tends to occur at the periphery, or boundaries, of the pathway, in particular at the gray-white matter junction, and within individual gyri.
Variability of shape features is influenced by scanner, vendor, acquisition, and b-value. Variability is shown as MAPE for each TractSeg, ATK, and Reco methods, for scan-rescan (RESCAN), scanners (SCAN1, SCAN2), vendor (VEN1, VEN2), resolution (RES1, RES2), diffusion directions (DIR1, DIR2) and b-value (BVAL). Values shown are averaged across all pathways within a bundle dissection method. Shape features are ordered (from top to bottom) from lowest to highest average MAPE. Many shape features are highly reproducible, and MAPE is influenced by all effects investigated.

Sources of variation may introduce bias in shape features. The mean percent variation (MPV), i.e., the signed MAPE, is shown for each bundle segmentation method, for all features, with the distribution across fiber pathways. A distribution not centered on 0 suggests systematic differences introduced by the given effect. For interpretation, RESCAN (repeat 2 – repeat 1), SCAN1 (Philips Achieva scanner 2 – Philips Achieva scanner 1), SCAN2 (Siemens Connectome standard acquisition – Siemens Prisma standard acquisition), VEN1 (GE Discovery - Philips Achieva), VEN2 (Siemens Skyra - Philips Achieva), RES1 (Prisma state-of-the-art 30 directions - Prisma standard acquisition), RES2 (Connectom state-of-the-art 30 directions - Connectom standard acquisition), DIR1 (Philips Achieva 96 directions – Philips Achieva 32 directions), DIR2 (Prisma state-of-the-art 60 directions - Prisma state-of-the-art 30 directions), BVAL (Prisma standard-acquisition b = 3000 - Prisma standard-acquisition b = 1000).

To look for systematic differences introduced in the quantification of features, we calculate the mean percent variation (i.e., the signed value of MAPE), across all sources of variation, for all features (across all bundles). Fig. 8 shows that most effects do not significantly bias bundle shape measures. For example, nearly all features derived from TractSeg are within a 10% variation and largely centered on 0. However, RES2 and VEN2 do introduce a small, but consistent, bias, in measures of surface area, end area, and volume (in this case, the higher resolution results in smaller values). Similarly, for ATK, a bias is observed in the opposite direction for the same features for effects of acquisition reso-
ditions. Additionally, b-value introduces a significant bias for ATK, with the higher b-value scan resulting in larger quantitative values for these features. Reco, in agreement with previous figures, has a much wider range of variation, and larger effects due to acquisition for features of Diameter, Surface Area, End Areas and Volume. Thus, different sources of variation may bias quantitative extraction of shape features, and bias them differently for different bundle segmentation methods.

3.5. Variation across bundle segmentation methods

Next, we compared the agreement of the same bundle, but across different bundle segmentation methods. Fig. 9 shows the Dice overlap for 14 common bundles, comparing each method to every other. There is a low-to-moderate agreement, with Dice overlap values between 0.1–0.5 for all pathways. In general, ATK was most similar to TractSeg and Reco for most bundles (with some exceptions), while Xtract was most dissimilar to all others. The AF, ILF, and MDLF, were the most dissimilar across methods.

Fig. 10 visualizes where agreement and disagreement occurs across bundle segmentation methods, with example-pathways AF and OR. Here, while most of the core agrees across methods, there is also a consistent disagreement across methods, particularly in the thickness of the bundle and in the regions of the temporal lobe for the AF and connections in the occipital lobe for the OR. Thus, instead of random differences due to noise, differences across methods are reproducible disagreement, likely caused by fundamental differences in the segmentation technique and structure to be segmented.

3.6. Variation in diffusion MRI microstructure measures

We next investigate reproducibility of microstructure measures due to the aforementioned sources of variation, and tractography variation.
Fig. 11 shows the MAPE of FA for all four bundle segmentation methods. In all cases, the standard-color boxplots are variations due to the queried source of variation alone, whereas the darker-shaded boxplots are due to the source of variation and the added variation of tractography variation. Most notably, the MAPE due to RESCAN, SCAN, VEN, DIR, and BVAL alone are highly similar for all segmentation methods, with only minor differences due to the slightly different representations of the pathways (Fig. 9). These results are in line with the literature, with variation <3% for SCAN rescan (Farrell et al., 2007; Landman et al., 2007; Wakana et al., 2007), with 5–15% due to scanner and vendor effects (L. Ning et al., 2020; CM Tax et al., 2019), and as much as 10% due to differences in acquisition and diffusion sensitization (Jones and Basser, 2004; L. Ning et al., 2020; Tax et al., 2020). Notably, the added variation due to tractography does indeed increase differences in FA (as indicated by a solid horizontal line) in many cases, although the% increase in variation is on average <5%.

Fig. 12 shows the MAPE of MD for different sources of variation. Most noticeable, MD is highly different when calculated using two different b-values, as expected (Novikov et al., 2018; Landman et al., 2007; Jones, 2004; De Luca et al., 2021; DK Jones et al., 1999), followed by differences due to vendors. Differences across RESCAN, SCAN, RES, and DIR are typically <5%. Again, the use of tractography adds to this variance, although on 3% or less on average.

4. Discussion

The primary focus of this work was to study variability of diffusion fiber tractography bundle segmentation, performing the same analysis on different datasets on different scanners or with different acquisition protocols. For the databases investigated here, we have shown that the process of tractography bundle segmentation shows significant variation across different acquisition resolution and across different vendors, with less, albeit significant, variation across scanners and across diffusion sensitization. Variation is indeed expected when scanning the same subject twice, with all other experimental parameters constant, due to imaging noise and the stochastic nature of the tractography process,
however, these additional sources of variation add potential confounds to tractography analysis that may bias measurements, limit aggregation of datasets, and hinder direct interpretation and meta-analysis of different results across studies. While the primary focus was on variation due to vendor and scanner effects, acquisition effects, and b-value effects, we also show the most bundle segmentation workflows are highly reproducible when running the same analysis on data acquired in different sessions, but with the same scanner and protocol.

It is well-known that microstructural features at different sites and with different protocols are not immediately comparable, and in fact significantly biased due to various effects. However, the process of tractography is largely dependent upon fiber orientation estimates, rather than features of the signal magnitude directly (i.e., MD/FA), and it is not immediately intuitive that differences in scanners, acquisitions, and b-values may lead to significantly different results. The results of this work suggest that, indeed, the results of tractography and across sites adds variability that must be considered in the interpretation of both microstructural and shape features of these pathways.

4.1. Do we need to harmonize tractography?

“Harmonization” can be considered any effort at reducing variability in quantitative metrics between different databases, scanners, and studies. We have known that the voxel-wise signal varies across sites, scanners, and acquisitions (as evidenced by the multitude of efforts in the literature to study effects on DTI-indices (Alexander et al., 2019; Jones et al., 2018; Novikov et al., 2018; Jeurissen et al., 2019; Raffelt et al., 2017; Chamberland et al., 2019; Yeatman et al., 2012; Maffei et al., 2019; Forkel et al., 2014; Hau et al., 2017; Hau et al., 2016; Sarubbo et al., 2019; Sarubbo et al., 2013; Neubert et al., 2015; Neubert et al., 2014)) and now confirm that the tractography process itself does as well, and have quantified the extent that tractography contributes to variability. The question becomes “do we need to harmonize tractography?”. The short answer is “yes”, the long answer is: harmonizing likely entails both harmonizing the signal (e.g., FA, MD, RISH measures), harmonizing orientation, reducing effects of resolution, and combining the strengths of different bundle segmentation approaches.

Fig. 12. Variation of MD. Effects of scan-rescan (RESCAN; blue), scanners (SCAN1, SCAN2; red), vendor (VEN1, VEN2; dark purple), resolution (RES1, RES2; pink), diffusion directions (DIR1, DIR2; green) and b-value (BVAL; light purple) on MAPE of the MD for all fiber bundles dissected using each technique. The left boxplots are indicative of the variability inherent to each effect, whereas the darker-hued (right) boxplots indicate the added variability due to differences in tractograms. For each, a Wilcoxon signed rank test is performed to investigate whether tractography adds to (or removes) significant variance to this metric, and statistical significance is indicated by a solid black line.
The field of diffusion MRI harmonization has grown in recent years, with significant efforts to make diffusion microstructural measures comparable across sites and scanners (Mirzaalian et al., 2016; L. Ning et al., 2020; Zhong et al., 2020; Cetin Karayumak et al., 2019; KM Huynh et al., 2019; Mirzaalian et al., 2018; Fortin et al., 2017). Yet, these endeavors have traditionally not considered variability of tractography, which is ultimately influenced at both the local scale of individual voxels and voxel-wise reconstruction as well as a global scale of connecting discrete orientation estimates across the brain.

It is unclear what “harmonizing” tractography may entail. Clearly, consistent orientation estimates are key, but also streamline generation algorithms robust to voxel-sizes, and also segmentation algorithms that are consistently able to identify streamlines belonging to a pathway-of-interest. With the vast array of options to reconstruct orientation, generate streamlines, and segment bundles, it may be impossible to harmonize data in a way that is appropriate for all methods. Some effort has been performed to harmonize fiber orientation estimation specifically across time or across scanners (Vishwesh Nath et al., 2018; Chen et al., 2016; Moyer et al., 2020; KM Huynh et al., 2019). It may be possible that harmonizing the microstructural measures themselves may remove some possible confounds (i.e., if FA is used as a stopping criteria). Similarly, it is possible that the application and process of tractography in a standard space (as performed for XTRACT), or at a standard resolution may remove confounds associated with image resolution. Alternatively, various multi-site methods used for scalar microstructure features, instead of harmonizing bundles of streamlines directly, may be utilized to harmonize features extracted from bundles. Finally, even while there is significant variation, large agreement occurs in the core of reconstructed white matter pathways, and weighting all derived measures and features by tract density, or isolating the trunk of the bundle (Yeotman et al., 2012), may remove sources of variation.

Reassuringly, the automated methods considered are fairly robust to these studied sources of variation. Visually, the pathways look remarkably similar across scanners, acquisition, and protocols (Fig. 2), for all methods. Quantitatively, methods such as TractSeg, which utilize orientation estimates alone, in combination with machine learning techniques in order to map out tract orientation maps, endpoints, and binary segmentations are highly reproducible. Similarly, the other methods, while quantitatively having moderately larger variation, show similar shapes, locations, and connectivity across all effects. A final possible harmonization approach may be to combine the strengths of the various algorithms, rethinking the process of bundle segmentation to possibly utilize some combination of machine learning (TractSeg), and a volume-based extraction prior to streamline generation, followed by atlas-based (ATK, Xtract), or shape-based filtering (Reco) in order to delineate bundles consistently across potential confounds.

4.2. Which confounds impact tractography the most?

It is important to emphasize that we are purposefully not attempting to “rank” algorithms, or suggest that ones are better than others. Even the methods with apparent lower reproducibility of features and shapes are still moderately robust, and different implementations of these algorithms may have yielded different quantitative values. For example, different thresholding could have been applied to both density-based (Xtract) or streamline-based (all others) methods to increase specificity (or vice-versa, specificity), or different whole-brain tractography could have been applied prior to bundle dissection using Recobundles. However, regardless of implementation and choices of hyperparameters, we expect methods to show similar dependencies to the investigated sources of variation.

To our knowledge, this is the first time that multiple sources of variation of tractography have been investigated together. Reproducibility across raters, across algorithms, and across scanners have previously been investigated. Our results allow comparison of the relative impact of changes across sites or scanners, and suggest that, in general, tial resolution leads to the most dramatic differences in resulting tractograms. Less tissue-based partial volume effects within the white matter may facilitate delineation of white matter bundles (F Rheault et al., 2020). Additionally, when quantifying volume overlap and shape features, voxel-wise partial volume effects may cause a higher (or lower) estimate due to the representation of the bundle as a binary volume at the given spatial resolution. Finally, orientation-based partial volume effects are observed with different spatial resolution (Jones et al., 2020; Schilling et al., 2017), leading to differences in accuracy of fiber orientation distributions, as well as fundamental differences in common diffusion measures such as FA (which are often used in the tracking process).

The second biggest contributor to variability was vendor differences. Differences across scanners are known to introduce variability due to factors of maximum gradient strength (and hence echo times and repetition times), field strengths, gradient nonlinearities, receive coil sensitivities, software version, and system calibration (Mirzaalian et al., 2016). Here, we show that differences in vendors are typically greater than that due to different scanners (yet same vendor) alone. Over and above scanner differences, vendors themselves may vary in algorithm choices, algorithms for acquisition, reconstruction, background noise reduction, multi-fusion (Griswold et al., 2002; Pruessmann et al., 1999), and pulse sequence implementation. Here, we have shown that in addition to inconsistencies in DTI measures across vendors consistently shown in previous studies (AK Prohl et al., 2019; VA Magnotta et al., 2012; Min et al., 2018) there is also a large inconsistency in tractography volumes and locations due to differences in vendors.

Reassuringly, variation of b-value and number of diffusion directions led to relatively consistent tractography. While it is well-known that angular resolution affects the ability to reconstruct fiber orientations (Jones et al., 2020; Schilling et al., 2018; Canales-Rodriguez et al., 2018; Tournier et al., 2013; Tournier et al., 2008; Prckovska et al., 2008), most reconstruction methods are robust with as few as 30 directions (or less). Similarly, while reconstruction algorithms are dependent on diffusion sensitization (Schilling et al., 2018; Deducci et al., 2014), the b-value did not significantly affect tractography results (although does affect quantitative metrics association with DTI).

It is also interesting that the relative magnitude of sources of variation depend on the bundle dissection method. While variability generally decreases from RESCAN, DIR, BVAL, SCAN, then VEN and RES, several notable exceptions occur. ATK is highly sensitive to the b-value. This is likely due to the fact that this automated tractography is reconstructed using Generalized Q-ball Imaging (Yeh et al., 2010), and tracking thresholds are determined by the normalized quantitative anisotropy, which is known to be highly dependent on b-value (Yeh et al., 2013). In contrast, XTRACT is a probabilistic method based largely on fiber orientation (and its dispersion) alone (from the ball-and-stick model (Sotiropoulos et al., 2012)), and different b-values give highly similar results of orientation (although dispersion will vary). XTRACT is also most sensitive to drastic change in resolution, likely caused by the probabilistic nature of the tractography process and subsequent thresholding for segmentation.

4.3. Shape variation and location of variation

This is to the best of our knowledge also the first time that reproducibility of different shape features of tractography has been investigated. While the variation across and within subjects has previously been studied (Yeh, 2020), it is important to understand cross-protocol and cross-scanner effects if these features are to be potential biomarkers in health and disease. These shape measures show similar patterns of variability, largest across resolution, vendors, and scanners, and smallest variation across repeats, directions, and b-values. More than variation, different resolutions and b-values can significantly bias measures, for example consistently overestimating volume and surface areas at lower resolutions where more partial volume effects are expected. De-
pending on tractography method, many features are remarkably robust, with MAPE below 5%, in line with that of microstructure features. We also investigated locations of differences and similarities by visualizing where there was consistent agreement and disagreement. Importantly, even with differences in acquisition and scanners, methods are able to consistently reproduce the major shape and location of the intended pathway, with differences most frequently occurring at the periphery, or edges, of the pathway, and along the white matter and gray matter interface. While features of shape and geometry may be biased due to sources of variation, these differences do not consistently occur at any one location or place along the pathway.

4.4. Different workflows

Over and above the typically studied sources of variation, we found that differences due to the choice of bundle segmentation workflows are most pronounced. For any given pathway, overlap from one workflow to another was low-to-moderate. This is in part due to the inherent sensitivity/specificity of different algorithms – for example Recobundles will look for clusters exhibiting a certain shape, while Tractseg is based on deep-learned segmentation, and Xtract will be highly dependent on the chosen threshold – but more importantly due to fundamental differences in how the pathway is dissected or defined (Schilling et al., 2020; Mandonnet et al., 2018). For example, the definition of a pathway by one method may be entirely different from another method, including choices in the presence or absence of connections to entire lobes or lobules, or differences in estimated spatial extent of pathways. While differences across methods were larger, they were importantly consistently different, meaning that comparing findings using different methods may result in differing conclusions on connectivity or microstructure. Differences between bundle segmentation workflows are also confounded by differences in the entire process of tractography, including differences in modeling, generation of streamlines (i.e., tractography), and bundle segmentation or filtering. Thus, it is intuitive that major differences exist when implementing different standard workflows to study the brain.

4.5. Microstructure variation

Finally, we looked at how much the variation in tractography contributes to the already existing cross-protocol and cross-scanner variation in dMRI measures. For FA, difference across scanners are known to be as much as 5–15% (L. Ning et al., 2020; CM Tax et al., 2019), and differences are expected due to different b-values, while scan-rescan reproducibility is high (<5%). The variation in tractography segmentation does indeed statistically significantly increase this variation for most effects, although the increase is typically very small and <5%. Similar results are observed for MD, although most changes are most pronounced for MD across different scanners. Thus, while tractography has the benefit of added specificity over simply propagating atlas-derived regions to subject-space, it does potentially increase variability in these measurements. Although methods such as tract-based spatial statistics (Smith et al., 2006) have been developed to mitigate these effects, we lose the added benefit of characterizing an index of interest along or within the full trajectory of the pathway.

4.6. Future studies and limitations

Future studies should investigate additional sources of variation. Manual dissection of fiber bundles gives the dissector the ability to interactively manipulate pathways to their liking (F Rheault et al., 2020), and it remains to be seen how this is influenced by scanner and site given the flexibility of this approach. Further, it is unknown whether these variabilities will matter in a clinical setting (Vanderweyden et al., 2020; Mancini et al., 2019; Fekonja et al., 2019; Essayed et al., 2017), although with the importance of determining pathway boundaries, we hypothesize that the partial volume effects due to acquisition resolution will possibly influence decision making. It is worth investigating the potentially large array of automated bundle segmentation methods that exist, as some are likely more/less appropriate when comparing or combining datasets with different confounds. Additionally, as alternative segmentation methods, or even whole-brain connectome analysis pipelines, are proposed, the use of open-source multi-site multi-subject datasets (Koller et al., 2021; Avesani et al., 2019; Jack et al., 2008) should be encouraged to investigate the successes and limitations of new approaches. Many algorithms for reconstruction and tractography are now able to utilize multiple diffusion shells, and the change in variability and precision of tractography using these techniques compared to isolated diffusion weightings should be compared, but is outside the scope of this work. As along-fiber quantification (Chamberland et al., 2019; Yeatman et al., 2012) has proven valuable in the research setting, it would be worthwhile to perform investigations which parallel the current study in order to ask how and where along the bundle differences occur due to different effects. This has been previously investigated, but is largely limited to scan-rescan analysis (Yeatman et al., 2012; Koller et al., 2021; Chandio et al., 2020), while the tract-averaged indices are still commonly utilized in neuroimaging studies.

A major limitation of the current study is the limited sample sizes of both datasets due to challenges associated with scanning the same subjects on different scanners and with different protocols. However, there are few multi-site multi-subject databases, and fewer still with varied protocols on the same subjects, whereas here we are able to remove effects across subjects by analyzing only the same subject with different protocols. It is expected that more datasets will become available as big-data and multi-site collaborations become more important to the neuroimaging community, and traveling subjects become common place in order to harmonize across sites. Exemplar open-source datasets include that of (Tong et al., 2020) with N = 3 subjects at 20 sites with Prisma scanners and a multi-shell dataset (allowing analysis of RESCAN, SCAN, BVAL, DIR), the traveling human phantom dataset (VA Magnotta et al., 2012) with N = 5 subjects at 8 center (SCAN, VEN, DIR), or consortia such as Pharmacoq (Galluzzi et al., 2016), ADNI (Jack et al., 2008), HCP (Glasser et al., 2013), or OASIS (Marcus et al., 2007), all with large sample size and repeat scans, but typically limited to RESCAN analysis only or without matched subjects across scanners/vendors/protocols. Because of this, for simplicity, we have chosen two datasets in this study which allow incorporation of all intended sources of variation without compromising readability. While we have looked at a wider range of variability factors than previous studies, we emphasize that these results are based only on two specific databases, and nalysis should be reproduced on other (and new) databases in future work to show generalizability.

Finally, while the primary focus of our study was on variation due to scanner-effects, acquisition-effects, and b-value-effects, our analysis was limited to studying these effects on only four bundle segmentation workflows. We did not implement all existing automated bundle reconstruction pipelines or workflows (Yeatman et al., 2012; Vazquez et al., 2020; F Zhang et al., 2019; Guevara et al., 2012; Wakana et al., 2007; Wassermann et al., 2019; F Zhang et al., 2020; F Zhang et al., 2019; O’Donnell and Westin, 2007; Wassermann et al., 2016; Ros et al., 2013; Zöllei et al., 2019; Schilling et al., 2021), however, our selection captures a variety of techniques used to reconstruction bundles, including differences in the use of atlases or regions-of-interest, those based on shape and/or orientation features, machine learning techniques, and differences in the generation of streamlines – a wide variety of vastly different approaches that we consider a strength of this study. To create a tractable parameter space, we have chosen only these four representatives of the wide variety of possible approaches. Finally, we did not directly perform harmonization techniques in this study. There are dozens of methods available to do this (see (L. Ning et al., 2020; CM Tax et al., 2019)), and understanding and characterizing harmonization results across several algorithms would take away from
the main focus of this study – which is characterization and ranking of variability across confounds. Further, harmonization would only affect a subset of results (i.e., those looking at FA/MD) as most harmonization approaches leave orientation untouched.

5. Conclusion

When investigating connectivity and microstructure of the white matter pathways of the brain using tractography, it is important to understand potential confounds and sources of variation in the process. Here, we find that tractography bundle segmentation results are influenced by the use of different vendors and scanners, and different acquisition choices of resolution, diffusion directions, and diffusion sensitizations, thus results may not be directly comparable when combining data or results across studies. Additionally, different bundle segmentation protocols have different successes/limitations when dealing with sources of variation, and the use of different protocols for bundle segmentation may result in different representations of the same intended pathway. These confounds need to be considered when designing or developing new tractography or bundle dissection algorithms, and when interpreting or combining data across sites.

6. Code

Multi-site, multi-scanner, multi-protocol, and multi-subject databases are available for MASIVar (https://openneuro.org/datasets/ds003416) and for MUSHAC (by request). Tractography pipelines are implemented as described by each software package using default parameters for TractSeg (Release 2.3; https://github.com/MIC-DKFZ/TractSeg), ATK (Lct 17 2020 build; http://dsi-studio.labsolver.org), RECO (Dipy 1.2.0 ; https://dipy.org), and XTRACT (FSL 6.0.3; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/XTRACT). Shape analysis is available in DSI Studio, as Matlab Code (https://github.com/dmitrishastin/tractography_shapes/).

Acknowledgements

This work was supported by the National Science Foundation Career Award #1452486, the National Institutes of Health under award numbers R01EB017230, and T32EB001628, and in part by VISE/VCIT VR3029 and the National Center for Research Resources, Grant UL1 RR024975-01. CMWT was supported by a Sir Henry Wellcome Fellowship (215944/Z/19/Z) and a Veni grant (17331) from the Dutch Research Council (NWO).

Data And Code Availability

The data for this study were selected from two open-sourced multi-subject, multi-scanner, and multi-protocol benchmark databases: the MASIVar (Alexander et al., 2019) and MUSHAC datasets (Jones et al., 2018; Novikov et al., 2018) – thus the data is freely available as found in associated references.


Credit Roles

KS: Conceptualization; Formal analysis; Investigation; Methodology; Writing;
CT: Conceptualization; Data curation; Writing
FR: Investigation; Methodology; Software
CH: Investigation; Methodology; Data curation
QY: Investigation; Methodology
FY: Conceptualization; Investigation; Methodology; Writing; Software
LC: Investigation; Methodology; Data curation
AA: Conceptualization; Methodology
Bl: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision

Appendix

The bundles resulting from each segmentation pipeline are given as a list below, with acronyms used in the text.

Recobundles:

Anterior Commissure (AC); Arcuate Fasciculus left (AF_L); Arcuate Fasciculus right (AF_R); Cerebellum left (CB_L); Cerebellum right (CB_R); Cingulum left (C_L); Cingulum right (C_R); Corpus Callosum (CC); Corticospinal Tract left (CST_L); Corticospinal Tract right (CST_R); Corticostriatal Pathway left (CS_L); Corticostriatal Pathway right (CS_R); Central Tegmental Tract left (CT_L); Central Tegmental Tract right (CT_R); Extreme Capsule left (EMC_L); Extreme Capsule right (EMC_R); Fornix left (F_L); Fornix right (F_R); Frontal Aslant Tract left (FAT_L); Frontal Aslant Tract right (FAT_R); Fronto-pontine tract left (FPT_L); Fronto-pontine tract right (FPT_R); Inferior Cerebellar Peduncle left (ICP_L); Inferior Cerebellar Peduncle right (ICP_R); Inferior Fronto-occipital Fasciculus left (IOFO_L); Inferior Fronto-occipital Fasciculus right (IOFO_R); Inferior Longitudinal Fasciculus left (ILF_L); Inferior Longitudinal Fasciculus right (ILF_R); Middle Cerebellar Peduncle (MCP); Middle Longitudinal Fasciculus left (MLF_L); Middle Longitudinal Fasciculus right (MLF_R); Medial Lemniscus left (ML_L); Medial Lemniscus right (ML_R); Occipito Pontine Tract left (OPT_L); Occipito Pontine Tract right (OPT_R); Optic Radiation left (OR_L); Optic Radiation right (OR_R); Parieto Pontine Tract left (PPT_L); Parieto Pontine Tract right (PPT_R); Superior Cerebellar Peduncle (SCP); Superior longitudinal fasciculus left (SLF_L); Superior longitudinal fasciculus right (SLF_R); Uncinate Fasciculus left (UF_L); Uncinate Fasciculus right (UF_R);
TractSeg:
Arcuate fasciculus left (AF_L); Arcuate fasciculus right (AF_R); Anterior Thalamic Radiation left (ATR_L); Thalamic Radiation right (ATR_R); Commissure Anterior (CA); Rostrum (CC_1);Genu (CC_2); Rostral body (Premotor) (CC_3); Anterior midbody (Primary Motor) (CC_4); Posterior midbody (Primary Somatosensory) (CC_5); Isthmus (CC_6); Splenium (CC_7); Corpus Callosum – all (CC); Cingulum left (CG_L); Cingulum right (CG_R); Corticospinal tract left (CST_L); Corticospinal tract right (CST_R); Fronto-pontine tract left (FPT_L); Fronto-pontine tract right (FPT_R); Fornix left (FX_L); Fornix right (FX_R); Inferior cerebellar peduncle left (ICP_L); Inferior cerebellar peduncle right (ICP_R); Inferior occipito-frontal fascicle left (IFO_L); Inferior occipito-frontal fascicle right (IFO_R); Inferior longitudinal fasciculus left (ILF_L); Inferior longitudinal fasciculus right (ILF_R); Middle cerebellar peduncle (MCP); Middle longitudinal fasciculus left (MLF_L); Middle longitudinal fasciculus right (MLF_R); Optic radiation left (OR_L); Optic radiation right (OR_R); Parieto-occipital pontine left (POPT_L); Parieto-occipital pontine right (POPT_R); Superior cerebellar peduncle left (SCP_L); Superior cerebellar peduncle right (SCP_R); Superior longitudinal fasciculus III left (SLF_III_L); Superior longitudinal fasciculus III right (SLF_III_R); Superior longitudinal fasciculus I left (SLF_I_L); Superior longitudinal fasciculus I right (SLF_I_R); Striato-fronto-orbital left (ST_FO_L); Striato-fronto-orbital right (ST_FO_R); Striato-occipital left (ST_OCC_L); Striato-occipital right (ST_OCC_R); Striato-parietal left (ST_PAR_L); Striato-parietal right (ST_PAR_R); Striato-postcentral left (ST_POSTC_L); Striato-postcentral right (ST_POSTC_R); Striato-precentral left (ST_PREC_L); Striato-precentral right (ST_PREC_R); Striato-perirolandic left (ST_PERI_L); Striato-perirolandic right (ST_PERI_R); Striato-posterior left (ST_POST_L); Striato-posterior right (ST_POST_R); Striato-posterior left (ST_POST_L); Striato-posterior right (ST_POST_R);
(ST_PREC_L); Striato-precentral right (ST_PREC_R); Striato-prefrontal left (ST_PREF_L); Striato-prefrontal right (ST_PREF_R); Striato-premotor left (ST_PREM_L); Striato-premotor right (ST_PREM_R); Thalamo-opercular left (T_OCC_L); Thalamo-opercular right (T_OCC_R); Thalamo-parietal left (T_PAR_L); Thalamo-parietal right (T_PAR_R); Thalamo-postcentral left (T_POSTC_L); Thalamo-postcentral right (T_POSTC_R); Thalamo-precentral left (T_PREC_L); Thalamo-precentral right (T_PREC_R); Thalamo-prefrontal left (T_PREF_L); Thalamo-prefrontal right (T_PREF_R); Thalamo-prefrontal left (T_PREM_L); Thalamo-prefrontal right (T_PREM_R); Uncinate fascicle left (UFL_L); Uncinate fascicle right (UF_R).

Xtract:
Anterior Commissure (AC); Arcuate Fascicle left (AF_L); Arcuate Fascicle right (AF_R); Auditory Radiation left (AR_L); Auditory Radiation right (AR_R); Anterior Thalamic Radiation left (ATR_L); Anterior Thalamic Radiation right (ATR_R); Cingulum Bundle Dorsal left (CBD_L); Cingulum Bundle Dorsal right (CBD_R); Cingulum Bundle Parahippocampal left (CBP_L); Cingulum Bundle Parahippocampal right (CBP_R); Cingulum Bundle Temporal left (CTBT_L); Cingulum Bundle Temporal right (CTBT_R); Corticospinal Tract left (CST_L); Corticospinal Tract right (CST_R); Frontal Aslant left (FA_L); Frontal Aslant right (FA_R); Forceps Major (FMA); Forceps Minor (FMF); Fornix left (FX_L); Fornix right (FX_R); Inferior Frontal-occipital Fasciculus left (IFO_L); Inferior Frontal-occipital Fasciculus right (IFO_R); Inferior Longitudinal Fasciculus left (ILF_L); Inferior Longitudinal Fasciculus right (ILF_R); Middle Cerebellar Peduncle (MCP); Medio-Dorsal Longitudinal Fasciculus left (MDLF_L); Medio-Dorsal Longitudinal Fasciculus right (MDLF_R); Optic Radiation left (OR_L); Optic Radiation right (OR_R); Superior Longitudinal Fasciculus 1 left (SLF1_L); Superior Longitudinal Fasciculus 1 right (SLF1_R); Superior Longitudinal Fasciculus 2 left (SLF2_L); Superior Longitudinal Fasciculus 2 right (SLF2_R); Superior Longitudinal Fasciculus 3 left (SLF3_L); Superior Longitudinal Fasciculus 3 right (SLF3_R); Superior Thalamic Radiation left (STR_L); Superior Thalamic Radiation right (STR_R); Uncinate Fasciculus left (UFL_L); Uncinate Fasciculus right (UF_R); Vertical Occipital Fasciculus left (VOFL_L); Vertical Occipital Fasciculus right (VOFR_R).

ATK:
Arcuate Fasciculus L (AF_L); Arcuate Fasciculus R (AF_R); Corticospinal Tract L (CST_L); Corticospinal Tract R (CST_R); Corticostriatal Pathway L (CSL_L); Corticostriatal Pathway R (CSR_R); Corticobulbar Tract L (CBT_L); Corticobulbar Tract R (CBT_R); Corticopontine Tract L (CPT_L); Corticopontine Tract R (CPT_R); Corticothalamic Pathway L (CPT_L); Corticothalamic Pathway R (CPT_R); Inferior Cerebellar Peduncle L (ICP_L); Inferior Cerebellar Peduncle R (ICP_R); Inferior Frontal Occipital Fasciculus left (IFO_L); Inferior Frontal Occipital Fasciculus right (IFO_R); Inferior Longitudinal Fasciculus left (ILF_L); Inferior Longitudinal Fasciculus right (ILF_R); Optic Radiation L (OR_L); Optic Radiation R (OR_R); Middle Longitudinal Fasciculus L (MLFL_L); Middle Longitudinal Fasciculus R (MLFL_R); Uncinate Fasciculus L (UF_L); Uncinate Fasciculus R (UF_R).

References
Jones, Huynh, Jones, Jones, Guevara, K.G., 2020. MRI (3), 10.1109/TMI.2019.2895020


