



HAL
open science

Tractostorm: Rater reproducibility assessment in tractography dissection of the pyramidal tract

Francois Rheault, Alessandro de Benedictis, Alessandro Daducci, Chiara Maffei, Chantal M W Tax, David Romascano, Eduardo Caverzasi, Felix C Morency, Francesco Corrivetti, Franco Pestilli, et al.

► **To cite this version:**

Francois Rheault, Alessandro de Benedictis, Alessandro Daducci, Chiara Maffei, Chantal M W Tax, et al.. Tractostorm: Rater reproducibility assessment in tractography dissection of the pyramidal tract. 2019. hal-03007052

HAL Id: hal-03007052

<https://hal.science/hal-03007052>

Preprint submitted on 16 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tractostorm: Rater reproducibility assessment in tractography dissection of the pyramidal tract

Francois Rheault, MSc^a, Alessandro De Benedictis, MD/PhD^b, Alessandro Daducci, PhD^c, Chiara Maffei, PhD^d, Chantal M.W Tax, PhD^e, David Romascano, MSc^f, Eduardo Caverzasi, MD/PhD^g, Felix C. Morency, MSc^h, Francesco Corrivetti, MDⁱ, Franco Pestilli, PhD^j, Gabriel Girard, PhD^f, Guillaume Theaud^a, Ilyess Zemmoura, MD/PhD^k, Janice Hau, PhD^l, Kelly Glavin^m, Keshi M. Jordan, PhD^g, Kristofer Pomiecko^m, Maxime Chamberland, PhD^e, Muhamed Barakovic, MSc^f, Nil Goyette^h, Philippe Poulin, MSc^a, Quentin Chenot, MScⁿ, Sandip S. Panesar, MD/MSc^o, Silvio Sarubbo, MD/PhD^p, Laurent Petit, PhD^q, Maxime Descoteaux, PhD^a

^aSherbrooke Connectivity Imaging Laboratory (SCIL), Université de Sherbrooke, Sherbrooke, Canada

^bNeurosurgery Unit, Department of Neuroscience and Neurorehabilitation, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy

^cComputer Science Department, University of Verona, Verona, Italy

^dAthinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, USA

^eCardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, Cardiff, United Kingdom

^fSignal Processing Lab (LTS5), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

^gDepartment of Neurology, University of California, San Francisco, USA

^hImeka, Sherbrooke, Canada

ⁱDépartement de neurochirurgie, Hôpital Lariboisière, Paris, France

^jDepartment of Psychological and Brain Sciences, Indiana University, Bloomington, USA

^kUMR 1253, iBrain, Université de Tours, Inserm, Tours, France

^lBrain Development Imaging Laboratories, Department of Psychology, San Diego State University, San Diego, California, USA

^mLearning Research & Development Center (LRDC), University of Pittsburgh, Pittsburgh, USA

ⁿISAE-SUPAERO, Toulouse, France

^oDepartment of Neurosurgery, Stanford University, Stanford, USA

^pDivision of Neurosurgery, Emergency Department, "S. Chiara" Hospital, Azienda Provinciale per i Servizi Sanitari (APSS), Trento, Italy

^qGroupe d'Imagerie Neurofonctionnelle, Institut des Maladies Neurodégénératives - UMR 5293, CNRS, CEA University of Bordeaux, Bordeaux, France

Abstract

Investigative studies of white matter (WM) brain structures using diffusion MRI (dMRI) tractography frequently require manual WM bundle segmentation, often called “*virtual dissection*”. Human errors and personal decisions make these manual segmentations hard to reproduce, which have not yet been quantified by the dMRI community. The contribution of this study is to provide the first large-scale, international, multi-center variability assessment of the “*virtual dissection*” of the pyramidal tract (PyT). Eleven (11) experts and thirteen (13) non-experts in neuroanatomy and “*virtual dissection*” were asked to perform 30 PyT segmentation and their results were compared using various voxel-wise and streamline-wise measures. Overall the voxel representation is always more reproducible

than streamlines ($\approx 70\%$ and $\approx 35\%$ overlap respectively) and distances between segmentations are also lower for voxel-wise than streamline-wise measures ($\approx 3\text{mm}$ and $\approx 6\text{mm}$ respectively). This needs to be seriously considered before using tract-based measures (e.g. bundle volume versus streamline count) for an analysis. We show and argue that future bundle segmentation protocols need to be designed to be more robust to human subjectivity. Coordinated efforts by the diffusion MRI tractography community are needed to quantify and account for reproducibility of WM bundle extraction techniques in this era of open and collaborative science.

Keywords: Diffusion MRI, White Matter, Tractography, Bundle segmentation, Intra-rater, inter-rater, Reproducibility

1 Introduction

DMRI tractography reconstructs streamlines modeling white matter (WM) connectivity. The set of all streamlines forms an object often called the *tractogram* [Jeurissen et al., 2017; Catani and De Schotten, 2008]. When specific hypotheses about known pathways, i.e. WM bundles, are investigated, neuroanatomists design “*dissection plans*” that contain anatomical landmarks and instructions to isolate the bundle of interest from this whole brain tractogram [Catani et al., 2002; Catani and De Schotten, 2008; Chenot et al., 2018; Hau et al., 2016]. Bundles can be segmented to study WM morphology, asymmetries, and then can be associated to specific functions [Lee Masson et al., 2017; Groeschel et al., 2014; Masson et al., 2018; Catani et al., 2007] with approaches similar to other brain structures [Lister and Barnes, 2009; Reitz et al., 2009]. Despite having similar anatomical definitions across publications, the absence of common segmentation protocols for tractography leads to differences that are for the most part unknown and unaccounted for. We need to know how variable our measurements are if we want to be able to have robust tract-based statistics in the future.

The need for a gold standard that quantifies human variability is well-known and well-studied in other fields, such as automatic image segmentation, cell counting or in machine learning [Kleesiek et al., 2016; Entis et al., 2012; Boccardi et al., 2011; Piccinini et al., 2014]. For applications such as hippocampi or tumor segmentation, thorough assessments of reproducibility and multiple iterations of manual segmentation protocols already exist [Boccardi et al., 2015; Frisoni et al., 2015]. These protocols were specifically designed to reduce the impact of human variability and help outcome comparison in large-scale clinical trials across multiple centers [Gwet, 2012; Frisoni et al., 2015].

The reproducibility of manual bundle segmentation will always be lower than manual image segmentation. Image segmentation in 3D requires local decision-making, and the decision to include voxels or not is directly done by raters. However, bundle segmentation requires local decisions that possibly impact the whole volume as streamlines reach outside of the scope of decisions made by raters. Since small hand-drawn regions of interest (ROI) or spheres are used to segment bundles, small mistakes can have far-reaching

*2500, boul. de l’Université, Sherbrooke (Québec) Canada, J1K 2R1

Email address: Francois.M.Rheault@USherbrooke.ca (Maxime Descoteaux, PhD)

30 consequences. Even if ROIs are fairly reproducible in a strict protocol, the resulting bun-
31 dles could be far from reproducible. This local-decision and global-impact conundrum
32 makes the design of reproducible protocols more difficult and can potentially cause low
33 agreement between raters.

34 1.1. Bundle segmentation

35 Bundle segmentation is the action of isolating streamlines based on neuroanatomical
36 priors, using known regions where certain conditions need to be satisfied. Inclusion and
37 exclusion regions-of-interests (ROIs) are drawn and defined at the voxel-level using co-
38 registered structural images, and are subsequently used to select the streamlines produced
39 by tractography [Catani et al., 2002; Behrens et al., 2007; Ghaziri et al., 2015; Renaud
40 et al., 2016; Rozanski et al., 2017], as seen in the Figure 1. Streamlines can be included
41 or discarded using inclusion ROIs where streamlines are forced to traverse, and exclusion
42 ROIs that cannot be crossed. Known structures such as grey nuclei, gyri or sulci and
43 recognizable signal signatures can be used as landmarks to create a plan to follow for the
44 segmentation [Catani et al., 2002; Catani and De Schotten, 2008; Hau et al., 2016; Chenot
45 et al., 2018]. In this work, the person performing the task of segmentation (i.e drawing
46 the ROIs, following the protocol) will be referred to as *rater*. Manual segmentation can
47 be performed in software such as, but not limited to, DTI studio [Jiang et al., 2006],
48 Trackvis [Wang et al., 2007], exploreDTI [Leemans et al., 2009], MITK Diffusion [Neher
49 et al., 2012], FiberNavigator [Chamberland et al., 2014], or MI-Brain [Rheault et al.,
50 2016] (Figure 1).

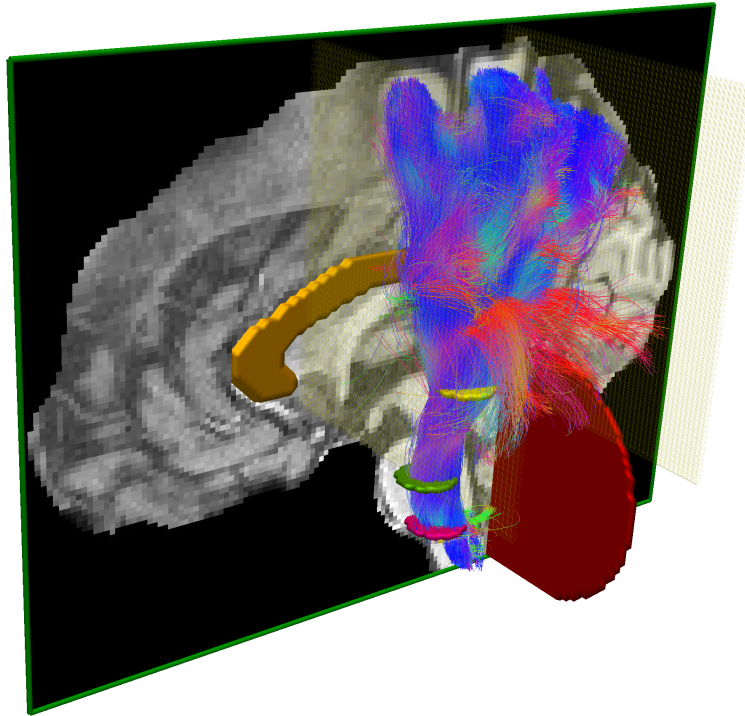


Figure 1: Illustration of the dissection plan of the PyT using the MI-Brain software [Rheault et al., 2016]. 3 axial inclusion ROIs (pink, green, yellow), 1 sagittal exclusion ROI (orange), 2 coronal exclusion ROIs (light yellow) and a cerebellum exclusion ROI (red). The whole brain tractogram was segmented to obtain the left pyramidal tract.

51 Once a bundle of interest is segmented from a tractogram, the analysis varies ac-
52 cording to the research question. It is common to report asymmetry or group difference
53 in bundle volume [Catani et al., 2007; Song et al., 2014; Chenot et al., 2018], diffusion
54 values within the bundle of interest (average fractional anisotropy, mean diffusivity, etc.)
55 [De Erausquin and Alba-Ferrara, 2013; Kimura-Ohba et al., 2016; Ling et al., 2012; Mole
56 et al., 2016] or values along the bundle (called *profilometry* and *tractometry*) [Dayan
57 et al., 2016; Yeatman et al., 2012, 2018; Cousineau et al., 2017]. Spatial distribution
58 of cortical terminations of streamlines can help to identify cortical regions with under-
59 lying WM connections affected by a condition [Rushworth et al., 2005; Johansen-Berg
60 et al., 2004; Donahue et al., 2016; Mars et al., 2011; Behrens et al., 2003]. Reporting
61 the number of streamlines (e.g streamline count in connectivity matrix or density maps)
62 is still very much present as a way to compare groups [Jones et al., 2013; Girard et al.,
63 2014; Sotiropoulos and Zalesky, 2017], despite not being directly related to anatomy or
64 connection strength [Jones, 2010; Jones et al., 2013].

65 *1.2. Quantifying reproducibility in tractography*

66 When performing segmentation, it is crucial that raters perform the tasks as closely
 67 as possible to the dissection plan. Even if a single individual performs all segmentations,
 68 the possibility of mistakes or erroneous decisions about landmarks exists [Boccardi et al.,
 69 2011; Frisoni et al., 2015; Entis et al., 2012]. High reproducibility is often an assump-
 70 tion, if this assumption is false the consequence could lead to inconsistent outcomes and
 71 erroneous conclusions. To assess the level of reproducibility of raters, identical datasets
 72 need to be segmented blindly more than once [Gisev et al., 2013; Gwet, 2012; Frisoni
 73 et al., 2015]. Reproducibility of segmentations from the same individual is referred to as
 74 intra-rater agreement, while reproducibility of segmentation across raters is referred to
 75 as inter-rater agreement.

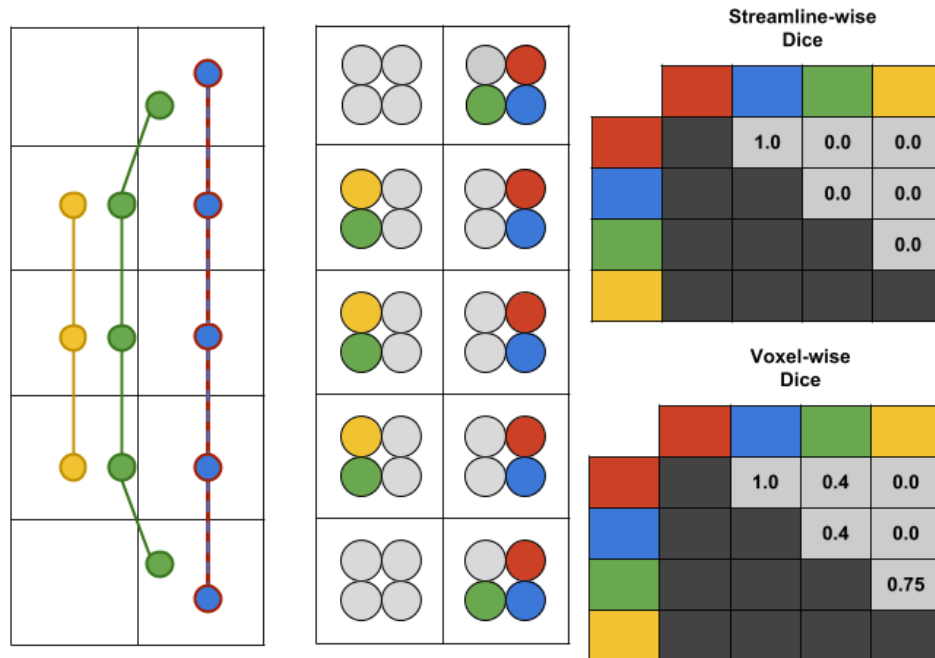


Figure 2: Representation of the Dice Coefficient (overlap) for both the streamline and the voxel representation. For the purpose of a didactic illustration, 4 streamlines are showed in a 2x5 voxel grid, the red and blue streamlines are identical. Each streamline is converted to a binary mask (point-based for simplicity) shown in a compact representation. Voxels with points from 3 different streamlines will results in voxels with 3 different colors, this can be seen as a spatial smoothing. The matrices on the right show values for all pairs (symmetrical). The green and yellow streamline are not identical, which results in a streamline-wise Dice coefficient of zero. However, in the voxel representation they have 3 voxels in common and the result is $(\frac{2*3}{5+3} = 0.75)$.

76 In the field of neuroimaging, voxels are used as the typical representation of data,
 77 while the available representation in tractography is in the form of streamlines (i.e. sets
 78 of 3D points in space). Figure 2 is a sketch of both representation. Several similarity
 79 measures exist to compare voxel-wise segmentations, e.g Dice Score. Most of them have

80 an equivalent formulation to compare sets of streamlines. However, resulting values can
81 widely vary as the spatial distribution is not the same for both representations. Some
82 measures related to streamlines require the datasets to be exactly the same, e.g Dice score,
83 as streamline reconstructions are sets of discrete points with floating point coordinates
84 and not discrete grids like 3D images. For this reason, comparison of streamlines is
85 more challenging and datasets that do not originate from the same source distance in
86 millimeters is often the only available solution [Garyfallidis et al., 2017; Maier-Hein et al.,
87 2017].

88 1.3. Summary of contributions of this work

89 Automatic segmentation methods are becoming more widespread [Guevara et al.,
90 2011; O'donnell et al., 2013; Chekir et al., 2014; Garyfallidis et al., 2017; Zhang et al.,
91 2018; Wasserthal et al., 2018] and aim to simplify the work of raters. The minimal
92 standard of any automatic segmentation method would be to reach the accuracy of
93 raters, thus it is crucial to truly quantify human reproducibility in manual tasks.

94 The goal of this work is first to quantify human reproducibility of bundle segmen-
95 tation from dMRI tractography. A measurement of rater (intra and inter) agreement is
96 extremely relevant to set an appropriate threshold for statistical significance. It is also
97 relevant for meta-analysis aiming to study large sets of publications and synthesize their
98 outcomes. An account of human errors or other sources of variability is necessary. A
99 second goal of this work is to investigate overlap, similarity measures and gold standard
100 comparison designed for tractography. Development of easily interpretable measures
101 for bundle comparison is necessary for large datasets. Overall the voxel representation
102 is significantly more reproducible than the streamline representation. The voxel rep-
103 resentation is better suited for analysis of tractography datasets (e.g reporting volume
104 instead of streamline count). More details about these different representations and
105 voxel/streamline-wise measures will be detailed in the [Method](#) and [Results](#) Section.

106 A thorough approach for bundle comparison quantification gives insights into seg-
107 mentation quality for future projects. This is needed to facilitate synthesis of findings
108 and outcomes from various publications [Gwet, 2012; Frisoni et al., 2015; Wisse et al.,
109 2017].

110 2. Method

111 2.1. Study design

112 Twenty-four participants were recruited and divided into two groups: experts and
113 non-experts. The division was based on their neuroanatomical educational background.
114 Participants working as researchers or PhD students in neuroanatomy, neurology or with
115 extended experience in the field performing “*virtual dissection*” as well as neurosurgeons
116 were part of the experts group (11 participants). The non-experts group was composed
117 of MSc, PhD student or Post-Doc in neuroimaging, but without any formal education
118 in neuroanatomy (13 participants). All participants had knowledge of dMRI tractogra-
119 phy in general as well as the concept of manual segmentations of tractography datasets.
120 Participation was voluntary and anonymous, recruitment was done individually and par-
121 ticipants from various labs in Europe and the USA were solicited. The study was per-
122 formed according to the guidelines of the Internal Review Board of the Centre Hospitalier
123 Universitaire de Sherbrooke (CHUS).

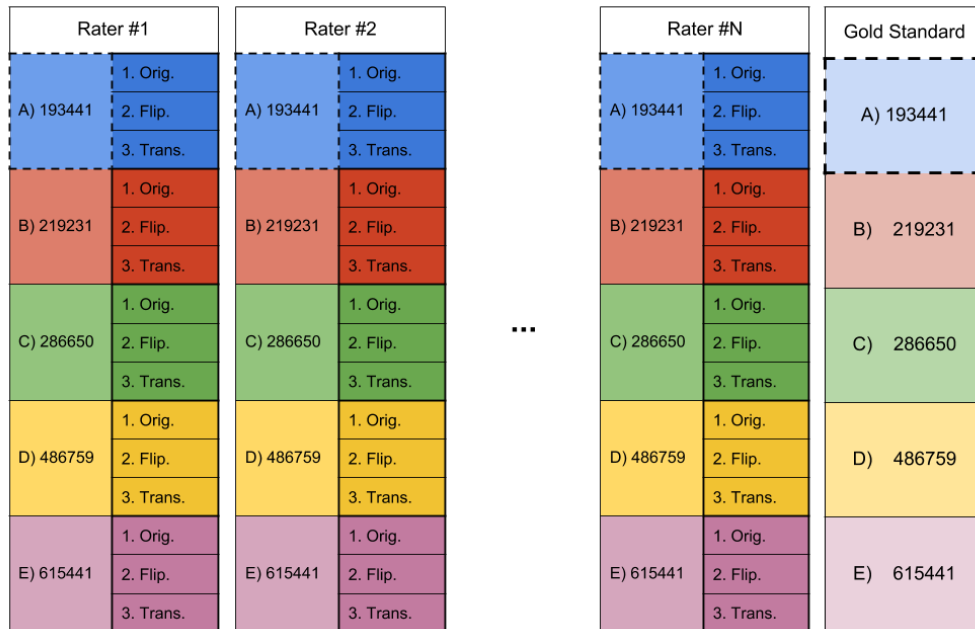


Figure 3: Representation of the study design showing N participants, each received 5 HCP datasets (listed and color-coded) which were replicated 3 times (original, flipped, translated). All participants had to perform the same dissection tasks, on the same anonymized datasets. Intra-rater, inter-rater and gold standard reproducibility were computed using the deanonymized datasets. More details are available in the supplementary materials

124 Five independent tractograms and their associated structural/diffusion images were
 125 used, each was triplicated (total of 15). One was untouched, one was flipped in the
 126 X axis (left/right) and one was translated. Then, all datasets were randomly named
 127 so the tasks could be performed blindly for each participant. Participants were not
 128 aware of the presence of duplicated datasets. Five tractograms were used to obtain
 129 stable averages, duplicated datasets were used to score the intra-rater agreement and
 130 the multiple participants to evaluate inter-rater agreement. The decision to separate
 131 participants in two groups was made to generate additional data about reproducibility
 132 in real-life conditions.

133 Figure 3 shows an overview of the study design. To evaluate intra-rater reproducibility
 134 of rater #1, each triplicate was used to compute reproducibility measures. Meaning that
 135 5 (A-B-C-D-E) x 3 (1-2-3) values were averaged to obtain the intra-rater “reproducibility
 136 score” of a single rater. To evaluate inter-rater reproducibility of rater #1, triplicates
 137 were fused and compared to all other raters to obtain a reproducibility measure. Meaning
 138 that 5 (A-B-C-D-E) x N (raters) values were averaged to obtain a single rater inter-rater
 139 “reproducibility score”. To evaluate reproducibility against the gold standard of rater #1
 140 the fused triplicates were also used. Meaning that 5 (A-B-C-D-E) x 1 (gold standard)
 141 values were averaged to obtain a single rater gold standard “reproducibility score”. The
 142 results showed in the Results Section are average values from all raters in each group.

143 All reproducibility measures were computed using the same approach.

144 *2.2. DWI datasets, processing and tractography*

145 Tractograms were generated from the preprocessed HCP [Van Essen et al., 2013]
146 DWI data using three shells (1000, 2000, 3000) with 270 directions. The B0, fractional
147 anisotropy (FA) and RGB (colored FA) images were computed from DWI to be used as
148 anatomical reference during segmentation. Constrained spherical deconvolution (CSD)
149 using a FA threshold from a tensor fit on the $b=1000s/mm^2$ was used to obtain fiber
150 orientation distribution functions (fODF) [Tournier et al., 2007; Descoteaux et al., 2007]
151 (spherical harmonic order 8) from the $b=2000s/mm^2$ and $b=3000s/mm^2$ shells. Prob-
152 abilistic particle filtering tractography [Girard et al., 2014] was subsequently computed
153 at 30 seeds per voxel in the WM mask (FSL FAST [Woolrich et al., 2009]) to make sure
154 sufficient density and spatial coverage were achieved.

155 The CSD model was also used for bundle-specific tractography (BST) to further
156 improve density and spatial coverage of the bundle of interest [Rheault et al., 2018;
157 Chenot et al., 2018]. This was to ensure that the full extent of the CST was reconstructed
158 and to ensure not to have criticisms from our experts in neuroanatomy complaining of
159 missing CST parts. A large model that approximates the CST was used to generate
160 streamlines with a strong preference for the Z axis (up-down). For BST, the same
161 tractography parameters were used except for seeding, which was exclusively done from
162 the precentral gyrus, postcentral gyrus and brainstem at 5 seeds per voxel.

163 The whole brain tractogram and the CST-specific tractogram were fused. To accom-
164 modate all participants and the wide range of computer performance, tractograms were
165 compressed using a 0.2mm tolerance error [Rheault et al., 2017; Presseau et al., 2015]
166 and commissural streamlines were removed and datasets split into hemispheres.

167 *2.3. Dissection plan and instructions*

168 Each participant received their randomly named datasets, a document containing
169 instructions for the segmentation and a general overview of a segmentation as example
170 (see supplementary materials). The segmentation task consisted in 15 segmentations of
171 the pyramidal tract (left and right). Segmentation involved using 3 WM inclusion ROIs
172 (Internal capsule, Midbrain and Medulla Oblongata) and 2 exclusion ROIs (one plane
173 anterior to the precentral gyrus and one plane posterior to the postcentral gyrus). The
174 detailed segmentation plan is available in the supplementary materials [Chenot et al.,
175 2018].

176 Participants had to perform the segmentation plans, following the instructions as
177 closely as possible. The dataset order was provided in a spreadsheet file. Participants had
178 to choose between two software; Trackvis [Wang et al., 2007] or MI-Brain [Rheault et al.,
179 2016]. This decision was made to guarantee that the data received from all participants
180 was compatible with the analysis.

181 Metadata such as date, starting time and duration had to be noted in the spreadsheet
182 file. Upon completion, the participants had to send back the same 15 folders with two
183 tractography files in each, the left and right pyramidal tract (PyT).

184 *2.4. Bundles analysis*

185 Once returned by all participants, datasets were de-randomized to match triplicates
186 across participants. The duplicates (flipped and translated) were reverted back to their
187 native space and all datasets (images and tractograms) were warped to a common space
188 (MNI152c 2009 nonlinear symmetrical) using the Ants registration library [Fonov et al.,
189 2011; Avants et al., 2008] to simplify the analysis. With all datasets having a uniform
190 naming convention and in a common space, the intra-rater and inter-rater reproducibility
191 can be assessed.

192 *Individual measures*

193 Reproducibility can be assessed using various measures. Volume and streamline count
194 are the main attributes obtained directly from files. They do not provide direct insight
195 about reproducibility, but one could expect that very similar segmentations should have
196 very similar values. However, this does not provide any nuance or specific information
197 about difference. In this work results for the left and right PyT are averaged together
198 without distinction, they are considered the same bundle during the analysis.

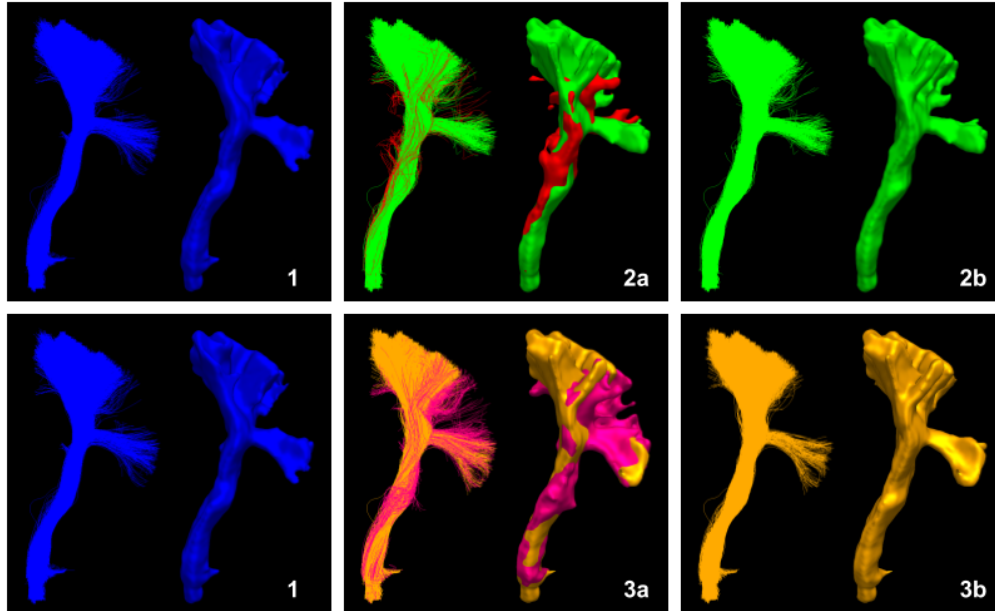


Figure 4: Comparison of bundles and the impacts of spurious streamlines on the reproducibility measurements. Each block shows streamlines on the left and the voxel representation on the right (isosurface). Block 2a and 3a shows the core (green/orange) and spurious (red/pink) portion of the bundle. Block 2b and 3b only shows the core portion of the bundle.

		1-2a	1-2b	1-3a	1-3b
Dice score	VOX	0.77	0.81	0.81	0.85
	STR	0.47	0.48	0.62	0.63
Bundle adjacency (mm)	VOX	2.66	2.64	2.04	1.82
	STR	4.41	3.54	4.63	3.24
Correlation of the density map	VOX	0.90	0.91	0.93	0.94

Figure 5: Table showing the reproducibility “score” between bundles, VOX marks voxel-wise measures and STR marks streamline-wise measures.

199 *Intra-rater and inter-rater*

200 Each participant performed the same tasks on each triplicate. The goal of this triplicate is to evaluate intra-rater reproducibility. Since all participants had access to the
 201 same datasets, inter-rater reproducibility can be assessed too.
 202

203 Computing the average value from all pairwise combinations provides an estimate of
 204 the agreement between multiple segmentations of a same bundle. The deviation can also
 205 provide insights about the consistency of these segmentations. Measurement values can
 206 be between 0 and 1, such as Dice and Jaccard [Dice, 1945], meaning they are independent
 207 of the size. An alternative to overlap measures are similarity measures, which can provide
 208 insights about the distance between two segmentations (in millimeter). Even when spatial
 209 overlap between two segmentations is low, both can be very similar in shape [Descoteaux
 210 et al., 2004; Garyfallidis et al., 2010]. Figure 5 shows bundles and how to interpret
 211 these measures. Pearson’s correlation coefficient obtained from density maps provides

212 insight into the statistical relationship and spatial agreement between two segmentations
213 [Hyde and Jesmanowicz, 2012]. More details on available measures for tractography are
214 available in the supplementary materials.

215 The most insightful measures are represented by the overlap (Dice coefficient), dis-
216 tance (bundle adjacency) and density and spatial coherence (density correlation). Each
217 measure provides a way to interpret the data at hand, but there is no single true measure
218 to summarize intra-rater and inter-rater agreement. Multiple measures were computed
219 and are all available in the supplementary materials along more detailed description for
220 each of them.

221 *Gold standard*

222 When multiple raters provide segmentations from an identical dataset, it is of interest
223 to produce a gold standard. For a voxel representation, a probability map can be con-
224 structed, where each voxel value represents the number of raters that counted the voxel
225 as part of their segmentation [Frisoni et al., 2015; Iglesias and Sabuncu, 2015; Langerak
226 et al., 2015; Pipitone et al., 2014]. This can be normalized and then thresholded to obtain
227 a binary mask representing whether or not the voxel was segmented by enough rater. A
228 threshold above 0.5 is often referred to as a majority vote. The same logic can be applied
229 to streamlines, each streamline can be assigned a value based on the number of raters
230 that considered it part of their segmentation.

231 This can be seen in Figure 6 where increasing the minimal vote threshold reduces the
232 number of outliers and overall size. In this work, the gold standard *does not* represent
233 the true anatomy and should not be interpreted as such. It simply represents the average
234 segmentation obtained from a tractogram.

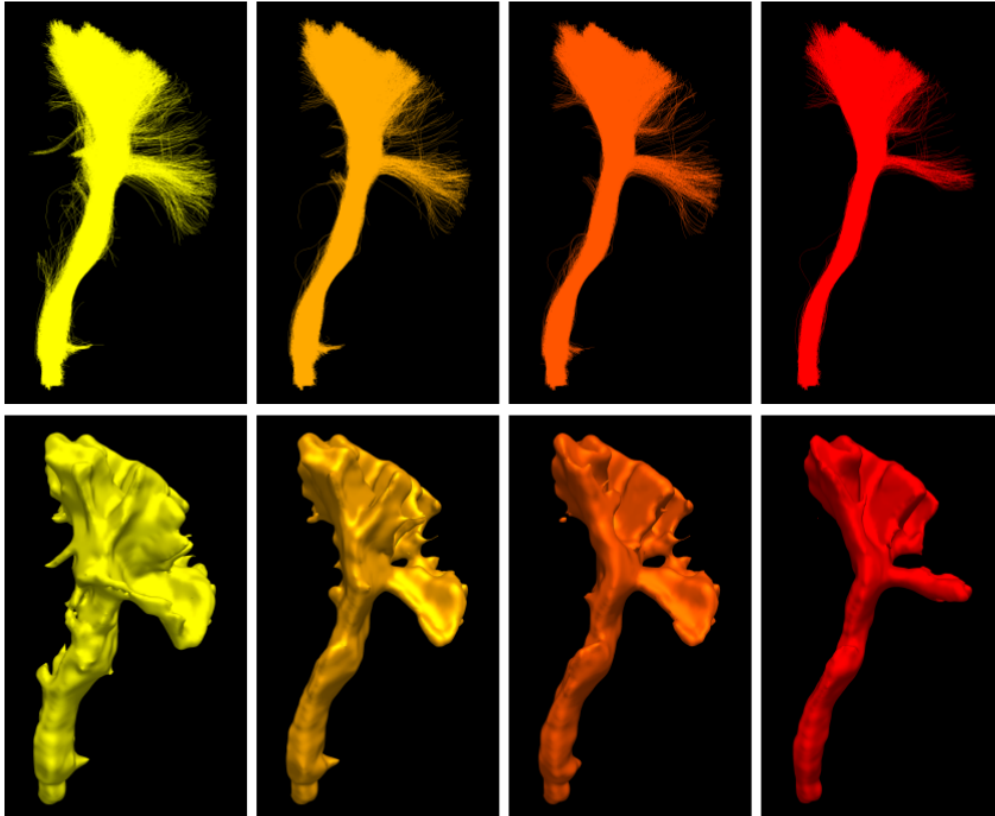


Figure 6: Gold standard obtained from 7 segmentations, first row shows the streamline representation and the second row shows the voxel represented as a smooth isosurface. From left to right, multiple voting ratios were used ($\frac{1}{7}, \frac{3}{7}, \frac{5}{7}, \frac{7}{7}$), each time reducing the number of streamlines and voxels consider part of the average segmentation. A minimal vote set at 1 out of 7 (left) is equivalent to a union of all segmentations while a vote set at 7 out of 7 (right) is equivalent to an intersection between all segmentations.

235 All elements that are not in a gold standard are true negatives and all the ones present
236 are true positives. By construction, the gold standard does not contain false positives
237 or false negatives. Binary classification measures are available such as sensitivity or
238 specificity. However, several other measures are available and each are a piece of the
239 puzzle leading to a more accurate interpretation [Garyfallidis et al., 2017; Chang et al.,
240 2009; Schilling et al., 2018].

241 To produce our gold standard a majority vote approach was used from the segmen-
242 tations of the experts group, as their knowledge of anatomy was needed to represent an
243 average version of the bundle of interest. The vote was set at 6 out of 11 and each of the
244 5 datasets got its own left and right gold standard. Since the representation at hand is
245 streamlines (which can be converted to voxels), a streamline-wise and a voxel-wise gold
246 standard were created.

247 3. Results

248 On average, experts produce “smaller” bundles than non-experts, their volume and
 249 streamline count is lower than non-experts, as it can be observed in Table 1 and Figure 7.
 250 This difference between groups is statically significant ($p - value < 0.01$). In the follow-
 251 ing sections, all explicit comparisons between groups are statistically significant using
 252 a standard Welch’s t-test for the means of two independent samples, which does not
 253 assume equal population variance ($p - value < 0.01$). The range of values for segmen-
 254 tation measures is wider for non-experts, meaning that either intra-rater or inter-rater
 255 variability is higher. As mentioned earlier, this is useful insight about reproducibility,
 256 but lacks nuance and context.

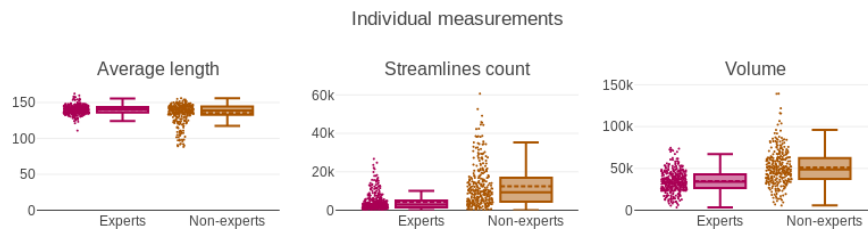


Figure 7: Boxplots and scatter plots showing distribution of the 3 measurements related to individual files for both groups.

		Expert		Non-experts	
		Mean	STD	Mean	STD
Volume (mm^3)	VOX	34835	12625	51146	20966
Streamline count	STR	4331	4457	12489	11091
Mean length (mm)	STR	140.33	7.81	138.70	11.29

Table 1: Table showing main values from boxplots of the 3 measurements related to individual files for both groups. The columns show the average value and the standard deviation for each group. VOX marks voxel-wise measures and STR marks streamline-wise measures. Rows shown in bold mean that the two groups (experts and non-experts) do not have the same distribution.

257 3.1. Intra-rater evaluation

258 All reported values can be seen in Table 2 and in Figure 8. The average intra-rater
 259 overlap is represented by the voxel-wise Dice coefficient and is on average 0.72 for experts
 260 and 0.78 for non-experts. Streamline-wise Dice coefficient is much lower at 0.31 and 0.52
 261 for both groups respectively. A higher Dice score value means that participants of a
 262 group are, on average, more reproducible with themselves. Non-overlapping voxels are
 263 on average at a distance 2.13mm for experts and 2.58mm for non-experts (lower Mean
 264 value represent higher similarity). Streamline-wise distance is lower in the experts group
 265 at 5.27mm while the non-experts group is at 6.12mm. The average density correlation is
 266 equal for both group at 0.82 and 0.82 for the experts and non-experts group respectively
 267 ($p - value > 0.01$).

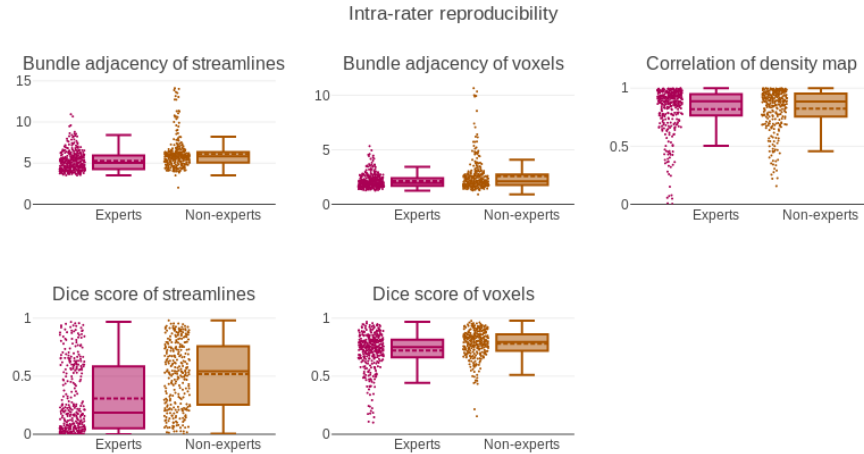


Figure 8: Boxplots and scatter plots showing distribution of the 3 measurements related to pairwise comparison measures for intra-rater segmentations.

		Expert		Non-experts	
		Mean	STD	Mean	STD
Dice score	VOX	0.72	0.16	0.78	0.12
	STR	0.31	0.30	0.52	0.28
Bundle adjacency (mm)	VOX	2.13	0.66	2.58	1.53
	STR	5.27	1.26	6.12	1.89
Correlation of density map	VOX	0.82	0.20	0.82	0.18

Table 2: Table showing main values from boxplots of the 3 measurements related to pairwise comparison measures for intra-rater segmentations. Voxel and streamline values of the same measures are in the same cell. Rows shown in bold mean that the two groups (experts and non-experts) do not have the same distribution.

268 3.2. Inter-rater evaluation

269 To minimize the influence of intra-rater reproducibility during the evaluation of inter-
 270 rater reproducibility, the triplicate datasets were fused into a single bundle. This was
 271 performed to approximate the results as if participant segmentations had no intra-rater
 272 variability. This lead to a underestimation of inter-rater variability, but necessary to
 273 separate source of variability later in the analysis. Voxel-wise Dice coefficient is on
 274 average higher between experts than between non-experts, at 0.75 and 0.67 respectively.
 275 Streamline-wise Dice coefficient is not statistically different ($p - value > 0.01$) at 0.34
 276 and 0.32. Voxel-wise distance is on average lower for the experts group than non-experts,
 277 2.74mm and 3.85mm respectively. The average density correlation is higher between
 278 experts at 0.88 while non-experts are at 0.71. The standard deviation is higher for the
 279 non-experts group, meaning that the similarity among non-experts is not only lower on
 280 average, but widely varies. All reported values can be seen in Table 3 and in Figure 9.

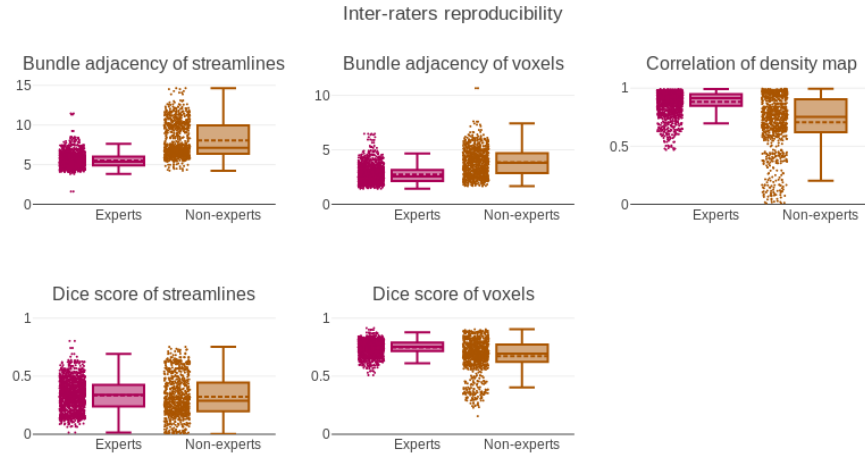


Figure 9: Boxplots and scatter plots showing distribution of the 3 measurements related to pairwise comparison measures for inter-rater segmentations.

		Expert		Non-experts	
		Mean	STD	Mean	STD
Dice score	VOX	0.75	0.06	0.67	0.14
	STR	0.34	0.13	0.32	0.18
Bundle adjacency (mm)	VOX	2.74	0.80	3.85	1.24
	STR	5.52	0.91	8.07	2.16
Correlation of density map	VOX	0.88	0.10	0.71	0.24

Table 3: Table showing main values from boxplots of the 3 measurements related to pairwise comparison measures for inter-rater segmentations. Voxel and streamline values of the same measures are in the same cell. Rows shown in bold mean that the two groups (experts and non-experts) do not have the same distribution.

281 3.3. Gold standard evaluation

282 All reported values can be seen in Table 4, 5 and in Figure 10, 11. Comparisons
 283 to the computed gold standard shows that on average experts and non-experts obtain
 284 segmentation roughly similar to the average segmentation. However, all measures show
 285 that segmentations from experts are on average closer to the gold standard than those of
 286 non-experts. This was expected as the gold standard was produced using segmentations
 287 from the experts group. Values for streamline-wise measures are lower for Dice coefficient
 288 and density correlation and higher for bundle adjacency, meaning that reproducibility is
 289 harder to achieve using the streamline representation. This was a similar trend observed
 290 in intra-rater and inter-rater values.

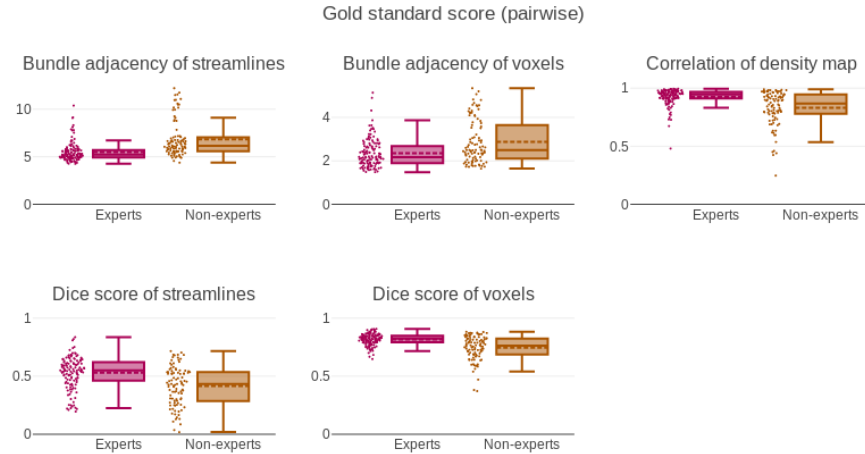


Figure 10: Boxplots and scatter plots showing distribution of the 3 measurements related to pairwise comparison measures against the gold standard.

		Expert		Non-experts	
		Mean	STD	Mean	STD
Dice score	VOX	0.82	0.05	0.74	0.10
	STR	0.53	0.14	0.42	0.17
Bundle adjacency (mm)	VOX	2.35	0.66	2.88	0.99
	STR	5.50	1.00	6.82	1.89
Correlation of density map	VOX	0.92	0.07	0.83	0.15

Table 4: Table showing main values from boxplots of the 3 measurements related to pairwise comparison measures against the gold standard. Voxel and streamline values of the same measures are in the same cell. Rows shown in bold mean that the two groups (experts and non-experts) do not have the same distribution.

291 Specificity and accuracy reach above the 95% for both groups both for streamlines
 292 or voxels. Meaning that experts and non-experts alike classified the vast majority of
 293 true negatives correctly. Since specificity is near a value of 1.0, the Youden score is
 294 almost equal to sensitivity. All 3 measures take into account the true negatives, which
 295 far outweigh the true positives, in our datasets, for this reason they were removed from
 296 Figure 11 and shown only in the supplementary materials. Sensitivity is much lower at
 297 0.59 and 0.71 for experts and non-experts respectively, as both groups partially capture
 298 the gold standard. Precision is higher for experts than for non-experts, meaning that
 299 experts were providing segmentations approximately the same *size* as the gold standard
 300 while non-experts were providing much bigger segmentations (that generally encompass
 301 the gold standard). This explains the higher sensitivity and lower specificity of non-
 302 experts. The average Kappa and Dice score is lower for experts at 0.67 and 0.72 while

303 the non-experts average is 0.69 and 0.73, respectively. The Kappa score takes into ac-
304 count overlap with the probability of randomly obtaining the right segmentation. Given
305 the dimensionality of our data, getting the right segmentation by accident is very low,
306 explaining why the Kappa and Dice score are very similar. It is important to consider
307 that the ratio of true negatives to true positives is not the same for both representations
308 (voxels vs. streamlines).

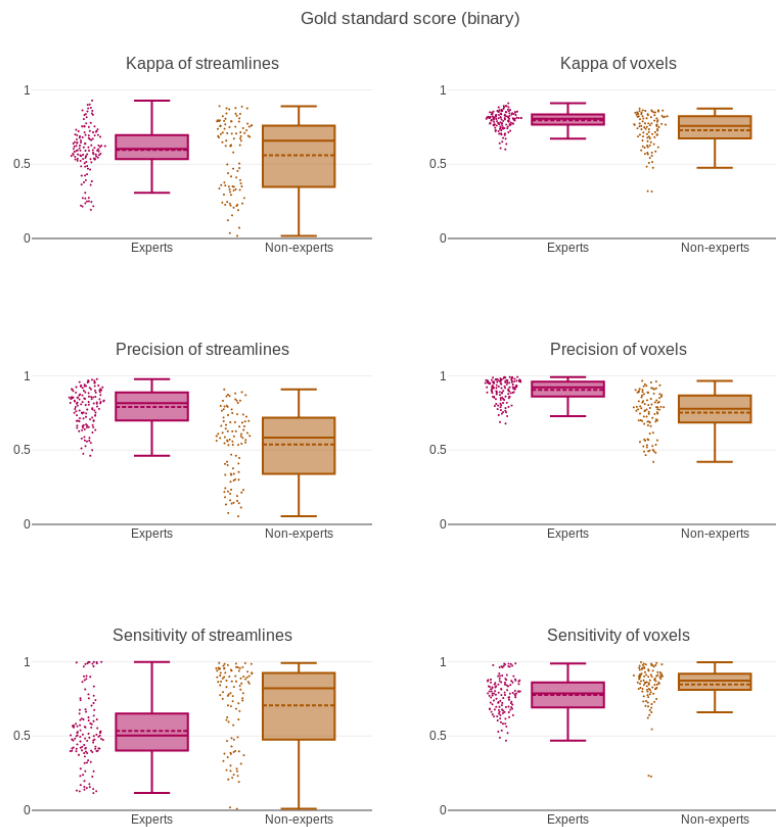


Figure 11: Boxplots and scatter plots showing distribution of the 6 measurements related to binary classification measures against the gold standard.

		Expert		Non-experts	
		Mean	STD	Mean	STD
Kappa	VOX	0.80	0.06	0.73	0.12
	STR	0.60	0.16	0.56	0.24
Precision	VOX	0.90	0.07	0.75	0.14
	STR	0.79	0.12	0.54	0.23
Sensitivity	VOX	0.78	0.12	0.85	0.13
	STR	0.54	0.23	0.71	0.27

Table 5: Table showing main values from boxplots of the 3 measurements related to binary classification measures against the gold standard. Rows shown in bold mean that the two groups (experts and non-experts) do not have the same distribution.

309 The computation of inter-rater reproducibility was performed using the fused tripli-
310 cate to minimize the influence of intra-rater reproducibility. The approach to fuse the
311 triplicate is simply an approximation, producing more than 3 segmentations of the same
312 datasets would be necessary to perfectly evaluate intra-rater reproducibility. However,
313 the 5 datasets used for this study represent sufficiently similar tasks to consider our ap-
314 proximation adequate for this work. Preliminary analysis showed low correlation values,
315 between a participant “score” for intra-rater reproducibility and inter-rater reproducibil-
316 ity. Correlation was between 0.2 and 0.4 for all measures, this indicates that there is
317 no clear link between the reproducibility of a participant’s own segmentations and the
318 agreement with other participants.

319 4. Discussion

320 4.1. Evaluation of protocols

321 This work illustrates that intra-rater and inter-rater agreement is far from perfect
322 even when following a strict and “simple” segmentation protocol. The intra-rater and
323 inter-rater agreement must be taken into account when researchers compare bundles ob-
324 tained from manual segmentations. When human expertise is required for a project, it
325 is crucial that people involved in the manual segmentation process evaluate their own
326 reproducibility, even if they have sufficient neuroanatomy knowledge and extensive expe-
327 rience in manual segmentation. This measure of error will likely increase the threshold
328 for statistical significance. In such case, either more datasets will be needed, or a better
329 protocol for segmentation needs to be designed [Gwet, 2012; Boccardi et al., 2015]. The
330 similarity between both groups indicates that with the right protocol, it would be pos-
331 sible to train people without anatomical background to perform tasks with results and
332 quality similar to experts.

333 Without such evaluation it is impossible for experts and non-experts alike to know
334 how reproducible they are beforehand. Since their “scores” vary with the protocol, the
335 bundle of interest and possibly other factors, it is important to consider an evaluation
336 before performing large-scale segmentation procedure [Frisoni et al., 2015]. An alternative
337 to guarantee more reproducible results is to design an appropriate protocol for non-
338 experts and to perform tasks blindly more than once. The results can then be averaged,
339 which will make outliers and errors easier to be identified.

340 This study did not allow for collaboration and did not micro-manage participants,
341 meaning they were left with the instructions without further intervention from the organ-
342 isers. In a situation where a segmentation plan can be defined in groups and techniques
343 can be improved along iterations of the plan, the intra-rater and inter-rater agreement
344 would likely go up. This study aimed at the evaluation of participants following in-
345 structions from a protocol, similar to the ones present in books, publications or online
346 examples.

347 *4.2. Measures and representations*

348 In this work the intra-rater agreement was higher for non-experts than experts, with-
349 out more information we could have concluded that non-experts were more meticulous
350 when they were performing their manual segmentations. However, by looking at sensi-
351 tivity and precision we can see that non-experts had “*bigger*” segmentations. Experts
352 are likely stricter in their decision-making process, this could amplify the local-decision
353 and global-impact conundrum mentioned earlier. A more liberal, less rigid, segmentation
354 likely makes it easier to be reproducible, but does not necessarily make it valid. This is
355 an example showing the importance of having more than one type of measure to obtain
356 a complete picture.

357 In tractography, it is common to use a single measure to portray a complex phe-
358 nomenon. Most measures used are simplified to have easily interpretable results. The
359 previous example shows the importance of using more than one type of measurements
360 to obtain a complete picture of the reproducibility. Reproducibility “*scores*” are likely
361 to vary with the project and the bundle of interest. This needs to be addressed as a
362 community. The discrepancy between protocol quality, reproducibility and conclusion
363 put forward in the literature can be problematic.

364 For binary measures (accuracy and specificity), scores were both above 95% as it
365 is easy to discard true negatives, and consequently did not provide much insight. Sim-
366 ilarly to the curse of dimensionality in machine learning [Verleysen and François, 2005;
367 Ceotto et al., 2011], our datasets typically contain millions of voxels (or streamlines), of
368 which only a few thousands true positives are considered during segmentation. Thus,
369 the vast majority of true negatives are rapidly discarded resulting in both accuracy and
370 specificity almost reaching 100%. Sensitivity provides more information, as true posi-
371 tives are more difficult to get, since they are rarer in the tractograms (few thousands
372 out of millions) [Maier-Hein et al., 2017]. This needs to be taken into account using
373 precision, as in some cases, strict segmentation is encouraged because false positives are
374 more problematic than false negatives. Streamline-wise measures show lower agreement,
375 meaning that reproducible results are likely more difficult to achieve with the streamlines
376 representation.

377 More complex measures need to be designed to fully capture the complexity of
378 tractography datasets and compare them, even across datasets or for longitudinal studies.
379 Currently, more advanced measures that capture fanning, spatial coherence, localized
380 curvature and torsion or spectral analysis are still rare, despite being used in other
381 neuroimaging fields [Esmaeil-Zadeh et al., 2010; Lombaert et al., 2012; Glzman et al.,
382 2018; Cheng and Basser, 2018].

383 4.3. Tractography algorithms

384 Iterative tractography algorithms are commonly divided in two categories: Deter-
385 ministic or probabilistic [Tournier et al., 2012; Garyfallidis et al., 2014]. The most striking
386 difference between both approaches is that probabilistic pathways cover more volume, as
387 they can easily curve and explore more ground. On the other hand, deterministic will
388 be more conservative due to curvature restrictions, thus leading to less exploration and
389 therefore smaller volume [Maier-Hein et al., 2017].

390 Manual segmentation of deterministic tractograms is likely more reproducible, since
391 small differences in ROI placement will have a smaller impact on the resulting bundle.
392 The local-decision and global-impact conundrum mentioned earlier is more obvious with
393 probabilistic tractography. Other tractography algorithms, such as global tractography
394 [Kreher et al., 2008; Mangin et al., 2013; Christiaens et al., 2015; Neher et al., 2012],
395 are likely to have different reproducibility “scores”, even with the same segmentation
396 protocol. Any change to the preprocessing could lead to unexpected change in the repro-
397 ducibility “scores”. Using the same datasets and tractography algorithm, but increasing
398 or decreasing the number of streamlines could also change the reproducibility “scores”.
399 Investigations of other bundles of interest would likely lead to different reproducibility
400 “scores”, using another anatomical definition of the PyT or even having the anatomi-
401 cal definition taught to participants would have the same effect. However, the general
402 conclusion remains that reproducibility needs to be quantified for specific projects and
403 protocols. Reproducibility “scores” cannot be generalized and any attempt would be
404 futile.

405 4.4. Impact on analysis

406 If variability needs to be minimized further than the defined protocol, a simple rec-
407 ommendation is to have a single rater perform each task multiple times or multiple raters
408 perform each task multiple times (or a subset of tasks). This way, it is guaranteed that
409 each dataset is segmented more than once, decreasing the error risk. Regardless of the
410 decision made, it is of great importance to quantify the reproducibility of manual segmen-
411 tation of raters involved in the project before doing any statistics or group comparisons.
412 This could drastically change the statistical significance of results. As of now, to the
413 best of our knowledge, human variability and errors are rarely considered. Sources of
414 variability needs to be accounted to truly enable synthesis of work across multiple cen-
415 ters. Even when automatic or semi-automatic methods are used, they first need to be
416 evaluated with agreed upon measures and reach or surpass human standards.

417 The extension to other bundles of interest or other segmentation plans is not trivial
418 and the only conclusion that stands is that agreement is never 100% and that a unique
419 measure is not enough to represent the whole picture for tractography segmentation. The
420 desire to simplify measures or have only one value to describe quality or reproducibility
421 of segmentations needs to be discouraged. The nature of our datasets makes this task
422 much more complex to interpret than 2D or 3D images, and it is imperative that the field
423 comes to understand and agree on measures to report. This is more relevant than ever as
424 the field grows and now that open science is becoming more popular and reproducibility
425 studies are encouraged. Similarly to other neuroimaging fields, such as hippocampi
426 segmentation, standardized protocols need to be developed and designed to be used
427 across multiple centers without active collaboration or micromanagement.

428 4.5. Future work

429 Future work includes the creation of a database containing bundle segmentations and
430 metadata from participants that will be available online so further analysis can be done.
431 As for now, a preliminary upload of the participants segmentation is available on Zenodo
432 (<https://doi.org/10.5281/zenodo.2547025>), which will be updated. In this work,
433 metadata was not used to evaluate duration as a variable influencing reproducibility.
434 Investigating the relationship between variability and duration of a task or looking for bias
435 (inter-hemispheric or software influence). An online platform similar to the Tractometer
436 [Côté et al., 2013] or a Nextflow pipeline [Di Tommaso et al., 2017] is planned to be
437 released. Such a tool would be designed for researchers to quickly submit data that is
438 expected to have some level of agreement and obtain their “*reproducibility score*”. This
439 way protocols can be improved and reproducibility can be taken into account in the
440 analysis.

441 Protocols for many bundles need to be developed for various purposes, such as
442 clinical practice, synthesis of findings, building training sets for machine learning, etc.
443 The segmentation plan and instructions need to be defined clearly by panels of experts,
444 and agreed upon terminology [Mandonnet et al., 2018], to optimize reproducibility and
445 anatomical validity. The field of manual tractography segmentation is decades behind
446 fields such as grey nuclei or hippocampi manual segmentation on this matter. The latter
447 has been refining segmentation protocols for the past decade and has already reached
448 the state harmonized segmentation protocol and was evaluated with reproducibility in
449 various settings [Boccardi et al., 2011, 2015; Frisoni et al., 2015; Apostolova et al., 2015;
450 Wisse et al., 2017].

451 5. Conclusion

452 When trying to understand how similar WM bundles from dMRI tractography are, at
453 least 3 values need to be taken into consideration: *Dice coefficient of voxels* showing how
454 well the overall volume overlaps, *bundle adjacency of voxels* showing how far are voxels
455 that do not overlap and *correlation of density map* showing if the streamlines are spatially
456 distributed in a similar way. Results from our work on the pyramidal tract revealed that
457 rater overlap is higher for voxel-wise measures (approximately 70%) than streamline-wise
458 measures (approximately 35%). Distance between segmentations is lower for voxel-wise
459 measures than streamline-wise measures, approximately 3mm and 6mm respectively. In
460 comparison to the group average, the results depict an ease to identify true negatives, an
461 adequate amount of true positives, while having a low amount of false positives. The voxel
462 and streamline representations do not produce equal levels of reproducibility. Studies
463 reporting bundle asymmetry in term of streamline count (streamline-based) will require
464 a larger group difference than those reporting volume difference (voxel-based). This
465 indicates a strong need for clear protocols for each bundle or at least detailed documents
466 included with publications that used manual segmentation. Reproducibility of results is
467 needed and goes hand-in-hand with the open science movement. A collaborative effort
468 to evaluate and quantify human variability is needed.

469 Acknowledgements

470 A special thanks to the funding sources for this work, the Fonds de recherche du
471 Québec - Nature et technologies (FRQNT) and Collaborative Research and Training
472 Experience Program in Medical Image Analysis (CREATE-MIA) programs. Thank you
473 to the Neuroinformatics Chair of the Sherbrooke University which helped push forward
474 neurosciences research.

475 References

- 476 Apostolova, L.G., Zarow, C., Biado, K., Hurtz, S., Boccardi, M., Somme, J., Honarpisheh, H., Blanken,
477 A.E., Brook, J., Tung, S., et al., 2015. Relationship between hippocampal atrophy and neuropathology
478 markers: a 7t mri validation study of the eadc-adni harmonized hippocampal segmentation protocol.
479 *Alzheimer's & Dementia* 11, 139–150.
- 480 Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration
481 with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical*
482 *image analysis* 12, 26–41.
- 483 Behrens, T.E., Berg, H.J., Jbabdi, S., Rushworth, M.F., Woolrich, M.W., 2007. Probabilistic diffusion
484 tractography with multiple fibre orientations: What can we gain? *Neuroimage* 34, 144–155.
- 485 Behrens, T.E., Johansen-Berg, H., Woolrich, M., Smith, S., Wheeler-Kingshott, C., Boulby, P., Barker,
486 G., Sillery, E., Sheehan, K., Ciccarelli, O., et al., 2003. Non-invasive mapping of connections between
487 human thalamus and cortex using diffusion imaging. *Nature neuroscience* 6, 750.
- 488 Boccardi, M., Bocchetta, M., Apostolova, L.G., Barnes, J., Bartzokis, G., Corbetta, G., DeCarli, C.,
489 Firbank, M., Ganzola, R., Gerritsen, L., et al., 2015. Delphi definition of the eadc-adni harmonized
490 protocol for hippocampal segmentation on magnetic resonance. *Alzheimer's & Dementia* 11, 126–138.
- 491 Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Cser-
492 nansky, J.G., De Leon, M.J., deToledo Morrell, L., et al., 2011. Survey of protocols for the manual
493 segmentation of the hippocampus: preparatory steps towards a joint eadc-adni harmonized protocol.
494 *Journal of Alzheimer's disease* 26, 61–75.
- 495 Catani, M., Allin, M.P., Husain, M., Pugliese, L., Mesulam, M.M., Murray, R.M., Jones, D.K., 2007.
496 Symmetries in human brain language pathways correlate with verbal recall. *Proceedings of the National*
497 *Academy of Sciences* 104, 17163–17168.
- 498 Catani, M., De Schotten, M.T., 2008. A diffusion tensor imaging tractography atlas for virtual in vivo
499 dissections. *cortex* 44, 1105–1132.
- 500 Catani, M., Howard, R.J., Pajevic, S., Jones, D.K., 2002. Virtual in vivo interactive dissection of white
501 matter fasciculi in the human brain. *Neuroimage* 17, 77–94.
- 502 Ceotto, M., Tantardini, G.F., Aspuru-Guzik, A., 2011. Fighting the curse of dimensionality in first-
503 principles semiclassical calculations: Non-local reference states for large number of dimensions. *The*
504 *Journal of chemical physics* 135, 214108.
- 505 Chamberland, M., Whittingstall, K., Fortin, D., Mathieu, D., Descoteaux, M., 2014. Real-time multi-
506 peak tractography for instantaneous connectivity display. *Frontiers in neuroinformatics* 8, 59.
- 507 Chang, H.H., Zhuang, A.H., Valentino, D.J., Chu, W.C., 2009. Performance measure characterization
508 for evaluating neuroimage segmentation algorithms. *Neuroimage* 47, 122–135.
- 509 Chekir, A., Descoteaux, M., Garyfallidis, E., Côté, M.A., Boumghar, F.O., 2014. A hybrid approach
510 for optimal automatic segmentation of white matter tracts in hardi, in: *Biomedical Engineering and*
511 *Sciences (IECBES), 2014 IEEE Conference on, IEEE*. pp. 177–180.
- 512 Cheng, J., Basser, P.J., 2018. Director field analysis (dfa): Exploring local white matter geometric
513 structure in diffusion mri. *Medical image analysis* 43, 112–128.
- 514 Chenot, Q., Tzourio-Mazoyer, N., Rheault, F., Descoteaux, M., Crivello, F., Zago, L., Mellet, E., Jobard,
515 G., Joliot, M., Mazoyer, B., et al., 2018. A population-based atlas of the human pyramidal tract in
516 410 healthy participants. *Brain Structure and Function* , 1–14.
- 517 Christiaens, D., Reisert, M., Dhollander, T., Sunaert, S., Suetens, P., Maes, F., 2015. Global tractog-
518 raphy of multi-shell diffusion-weighted imaging data using a multi-tissue model. *Neuroimage* 123,
519 89–101.
- 520 Côté, M.A., Girard, G., Boré, A., Garyfallidis, E., Houde, J.C., Descoteaux, M., 2013. Tractometer:
521 towards validation of tractography pipelines. *Medical image analysis* 17, 844–857.

- 522 Cousineau, M., Jodoin, P.M., Garyfallidis, E., Côté, M.A., Morency, F.C., Rozanski, V., GrandMaison,
523 M., Bedell, B.J., Descoteaux, M., 2017. A test-retest study on parkinson's ppmi dataset yields
524 statistically significant white matter fascicles. *NeuroImage: Clinical* 16, 222–233.
- 525 Dayan, M., Monohan, E., Pandya, S., Kuceyeski, A., Nguyen, T.D., Raj, A., Gauthier, S.A., 2016.
526 Profilometry: a new statistical framework for the characterization of white matter pathways, with
527 application to multiple sclerosis. *Human brain mapping* 37, 989–1004.
- 528 De Erausquin, G.A., Alba-Ferrara, L., 2013. What does anisotropy measure? insights from increased and
529 decreased anisotropy in selective fiber tracts in schizophrenia. *Frontiers in integrative neuroscience* 7,
530 9.
- 531 Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R., 2007. Regularized, fast, and robust analytical
532 q-ball imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for*
533 *Magnetic Resonance in Medicine* 58, 497–510.
- 534 Descoteaux, M., Collins, L., Siddiqi, K., 2004. Geometric flows for segmenting vasculature in mri: Theory
535 and validation, in: *International Conference on Medical Image Computing and Computer-Assisted*
536 *Intervention*, Springer. pp. 500–507.
- 537 Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., Notredame, C., 2017. Nextflow
538 enables reproducible computational workflows. *Nature biotechnology* 35, 316–319.
- 539 Dice, L.R., 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 297–302.
- 540 Donahue, C.J., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Behrens, T.E., Dyrby, T.B.,
541 Coalson, T., Kennedy, H., Knoblauch, K., Van Essen, D.C., et al., 2016. Using diffusion tractography
542 to predict cortical connection strength and distance: a quantitative comparison with tracers in the
543 monkey. *Journal of Neuroscience* 36, 6758–6770.
- 544 Entis, J.J., Doerga, P., Barrett, L.F., Dickerson, B.C., 2012. A reliable protocol for the manual segmen-
545 tation of the human amygdala and its subregions using ultra-high resolution mri. *Neuroimage* 60,
546 1226–1235.
- 547 Esmaeil-Zadeh, M., Soltanian-Zadeh, H., Jafari-Khouzani, K., 2010. Spharm-based shape analysis of
548 hippocampus for lateralization in mesial temporal lobe epilepsy, in: *Electrical Engineering (ICEE),*
549 *2010 18th Iranian Conference on, IEEE*. pp. 39–44.
- 550 Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C., et al.,
551 2011. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* 54, 313–327.
- 552 Frisoni, G.B., Jack Jr, C.R., Bocchetta, M., Bauer, C., Frederiksen, K.S., Liu, Y., Preboske, G., Swihart,
553 T., Blair, M., Cavado, E., et al., 2015. The eadc-adni harmonized protocol for manual hippocampal
554 segmentation on magnetic resonance: evidence of validity. *Alzheimer's & Dementia* 11, 111–125.
- 555 Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., Van Der Walt, S., Descoteaux, M., Nimmo-
556 Smith, I., 2014. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics* 8,
557 8.
- 558 Garyfallidis, E., Brett, M., Nimmo-Smith, I., 2010. Fast dimensionality reduction for brain tractography
559 clustering, in: *16th Annual Meeting of the Organization for Human Brain Mapping*.
- 560 Garyfallidis, E., Côté, M.A., Rheault, F., Sidhu, J., Hau, J., Petit, L., Fortin, D., Cunanne, S., De-
561 scoteaux, M., 2017. Recognition of white matter bundles using local and global streamline-based
562 registration and clustering. *NeuroImage* .
- 563 Ghaziri, J., Tucholka, A., Girard, G., Houde, J.C., Boucher, O., Gilbert, G., Descoteaux, M., Lippé, S.,
564 Rainville, P., Nguyen, D.K., 2015. The corticocortical structural connectivity of the human insula.
565 *Cerebral cortex* 27, 1216–1228.
- 566 Girard, G., Whittingstall, K., Deriche, R., Descoteaux, M., 2014. Towards quantitative connectivity
567 analysis: reducing tractography biases. *Neuroimage* 98, 266–278.
- 568 Gisev, N., Bell, J.S., Chen, T.F., 2013. Interrater agreement and interrater reliability: key concepts,
569 approaches, and applications. *Research in Social and Administrative Pharmacy* 9, 330–338.
- 570 Glozman, T., Bruckert, L., Pestilli, F., Yecies, D.W., Guibas, L.J., Yeom, K.W., 2018. Framework for
571 shape analysis of white matter fiber bundles. *NeuroImage* 167, 466–477.
- 572 Groeschel, S., Tournier, J.D., Northam, G.B., Baldeweg, T., Wyatt, J., Vollmer, B., Connelly, A., 2014.
573 Identification and interpretation of microstructural abnormalities in motor pathways in adolescents
574 born preterm. *NeuroImage* 87, 209–219.
- 575 Guevara, P., Poupon, C., Rivière, D., Cointepas, Y., Descoteaux, M., Thirion, B., Mangin, J.F., 2011.
576 Robust clustering of massive tractography datasets. *NeuroImage* 54, 1975–1993.
- 577 Gwet, K.L., 2012. Handbook of inter-rater reliability: The definitive guide to measuring the extent of
578 agreement among multiple raters. *Advanced Analytics, LLC* .
- 579 Hau, J., Sarubbo, S., Perchey, G., Crivello, F., Zago, L., Mellet, E., Jobard, G., Joliot, M., Mazoyer,
580 B.M., Tzourio-Mazoyer, N., et al., 2016. Cortical terminations of the inferior fronto-occipital and

- 581 uncinate fasciculi: anatomical stem-based virtual dissection. *Frontiers in neuroanatomy* 10, 58.
582 Hyde, J.S., Jesmanowicz, A., 2012. Cross-correlation: an fmri signal-processing strategy. *NeuroImage*
583 62, 848–851.
- 584 Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. *Medical*
585 *image analysis* 24, 205–219.
- 586 Jeurissen, B., Descoteaux, M., Mori, S., Leemans, A., 2017. Diffusion mri fiber tractography of the
587 brain. *NMR in Biomedicine* .
- 588 Jiang, H., Van Zijl, P.C., Kim, J., Pearlson, G.D., Mori, S., 2006. Dtistudio: resource program for diffu-
589 sion tensor computation and fiber bundle tracking. *Computer methods and programs in biomedicine*
590 81, 106–116.
- 591 Johansen-Berg, H., Behrens, T., Robson, M., Drobnjak, I., Rushworth, M., Brady, J., Smith, S., Higham,
592 D., Matthews, P., 2004. Changes in connectivity profiles define functionally distinct regions in human
593 medial frontal cortex. *Proceedings of the National Academy of Sciences* 101, 13335–13340.
- 594 Jones, D.K., 2010. Challenges and limitations of quantifying brain connectivity in vivo with diffusion
595 mri. *Imaging in Medicine* 2, 341.
- 596 Jones, D.K., Knösche, T.R., Turner, R., 2013. White matter integrity, fiber count, and other fallacies:
597 the do's and don'ts of diffusion mri. *Neuroimage* 73, 239–254.
- 598 Kimura-Ohba, S., Yang, Y., Thompson, J., Kimura, T., Salayandia, V.M., Cosse, M., Yang, Y., Sillerud,
599 L.O., Rosenberg, G.A., 2016. Transient increase of fractional anisotropy in reversible vasogenic edema.
600 *Journal of Cerebral Blood Flow & Metabolism* 36, 1731–1743.
- 601 Kleesiek, J., Petersen, J., Döring, M., Maier-Hein, K., Köthe, U., Wick, W., Hamprecht, F.A., Bendszus,
602 M., Biller, A., 2016. Virtual raters for reproducible and objective assessments in radiology. *Scientific*
603 *reports* 6, 25007.
- 604 Kreher, B., Mader, I., Kiselev, V., 2008. Gibbs tracking: a novel approach for the reconstruction of
605 neuronal pathways. *Magnetic Resonance in Medicine* 60, 953–963.
- 606 Langerak, T.R., van der Heide, U.A., Kotte, A.N., Berendsen, F.F., Pluim, J.P., 2015. Improving
607 label fusion in multi-atlas based segmentation by locally combining atlas selection and performance
608 estimation. *Computer Vision and Image Understanding* 130, 71–79.
- 609 Lee Masson, H., Wallraven, C., Petit, L., 2017. can touch this: Cross-modal shape categorization
610 performance is associated with microstructural characteristics of white matter association pathways.
611 *Human brain mapping* 38, 842–854.
- 612 Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. Exploredti: a graphical toolbox for processing,
613 analyzing, and visualizing diffusion mr data, in: 17th annual meeting of intl soc mag reson med,
614 International Society for Magnetic Resonance in Medicine Berkeley, CA, USA. p. 3537.
- 615 Ling, J.M., Pena, A., Yeo, R.A., Merideth, F.L., Klimaj, S., Gasparovic, C., Mayer, A.R., 2012. Biomark-
616 ers of increased diffusion anisotropy in semi-acute mild traumatic brain injury: a longitudinal per-
617 spective. *Brain* 135, 1281–1292.
- 618 Lister, J.P., Barnes, C.A., 2009. Neurobiological changes in the hippocampus during normative aging.
619 *Archives of Neurology* 66, 829–833.
- 620 Lombaert, H., Grady, L., Polimeni, J.R., Cheriet, F., 2012. Focusr: Feature oriented correspondence
621 using spectral regularization-a method for accurate surface matching. *IEEE transactions on pattern*
622 *analysis and machine intelligence* , 1.
- 623 Maier-Hein, K.H., Neher, P.F., Houde, J.C., Côté, M.A., Garyfallidis, E., Zhong, J., Chamberland, M.,
624 Yeh, F.C., Lin, Y.C., Ji, Q., et al., 2017. The challenge of mapping the human connectome based on
625 diffusion tractography. *Nature communications* 8, 1349.
- 626 Mandonnet, E., Sarubbo, S., Petit, L., 2018. The nomenclature of human white matter association
627 pathways: Proposal for a systematic taxonomic anatomical classification. *Frontiers in Neuroanatomy*
628 12, 94.
- 629 Mangin, J.F., Fillard, P., Cointepas, Y., Le Bihan, D., Frouin, V., Poupon, C., 2013. Toward global
630 tractography. *Neuroimage* 80, 290–296.
- 631 Mars, R.B., Jbabdi, S., Sallet, J., O'Reilly, J.X., Croxson, P.L., Olivier, E., Noonan, M.P., Bergmann,
632 C., Mitchell, A.S., Baxter, M.G., et al., 2011. Diffusion-weighted imaging tractography-based parcel-
633 lation of the human parietal cortex and comparison with human and macaque resting-state functional
634 connectivity. *Journal of Neuroscience* 31, 4087–4100.
- 635 Masson, H.L., Kang, H.m., Petit, L., Wallraven, C., 2018. Neuroanatomical correlates of haptic object
636 processing: combined evidence from tractography and functional neuroimaging. *Brain Structure and*
637 *Function* 223, 619–633.
- 638 Mole, J.P., Subramanian, L., Bracht, T., Morris, H., Metzler-Baddeley, C., Linden, D.E., 2016. Increased
639 fractional anisotropy in the motor tracts of parkinson's disease suggests compensatory neuroplasticity

- 640 or selective neurodegeneration. *European radiology* 26, 3327–3335.
- 641 Neher, P.F., Stieltjes, B., Reisert, M., Reicht, I., Meinzer, H.P., Fritzsche, K.H., 2012. Mitk global
642 tractography, in: *Medical Imaging 2012: Image Processing*, International Society for Optics and
643 Photonics. p. 83144D.
- 644 O'donnell, L.J., Golby, A.J., Westin, C.F., 2013. Fiber clustering versus the parcellation-based connec-
645 tome. *NeuroImage* 80, 283–289.
- 646 Piccinini, F., Tesei, A., Paganelli, G., Zoli, W., Bevilacqua, A., 2014. Improving reliability of live/dead
647 cell counting through automated image mosaicing. *Computer methods and programs in biomedicine*
648 117, 448–463.
- 649 Pipitone, J., Park, M.T.M., Winterburn, J., Lett, T.A., Lerch, J.P., Pruessner, J.C., Lepage, M.,
650 Voineskos, A.N., Chakravarty, M.M., Initiative, A.D.N., et al., 2014. Multi-atlas segmentation of
651 the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage*
652 101, 494–512.
- 653 Presseau, C., Jodoin, P.M., Houde, J.C., Descoteaux, M., 2015. A new compression format for fiber
654 tracking datasets. *NeuroImage* 109, 73–83.
- 655 Reitz, C., Brickman, A.M., Brown, T.R., Manly, J., DeCarli, C., Small, S.A., Mayeux, R., 2009. Linking
656 hippocampal structure and function to memory performance in an aging population. *Archives of*
657 *neurology* 66, 1385–1392.
- 658 Renaud, E., Descoteaux, M., Bernier, M., Garyfallidis, E., Whittingstall, K., 2016. Semi-automatic
659 segmentation of optic radiations and lgn, and their relationship to eeg alpha waves. *PloS one* 11,
660 e0156436.
- 661 Rheault, F., Houde, J.C., Descoteaux, M., 2017. Visualization, interaction and tractometry: Dealing
662 with millions of streamlines from diffusion mri tractography. *Frontiers in neuroinformatics* 11, 42.
- 663 Rheault, F., Houde, J.C., Goyette, N., Morency, F., Descoteaux, M., 2016. Mi-brain, a software to handle
664 tractograms and perform interactive virtual dissection, in: *ISMRM Diffusion study group workshop*,
665 Lisbon.
- 666 Rheault, F., St-Onge, E., Sidhu, J., Chenot, Q., Petit, L., Descoteaux, M., 2018. Bundle-specific
667 tractography, in: *Computational Diffusion MRI*. Springer, pp. 129–139.
- 668 Rozanski, V.E., da Silva, N.M., Ahmadi, S.A., Mehrkens, J., da Silva Cunha, J., Houde, J.C., Vollmar,
669 C., Bötzel, K., Descoteaux, M., 2017. The role of the pallidothalamic fibre tracts in deep brain
670 stimulation for dystonia: a diffusion mri tractography study. *Human brain mapping* 38, 1224–1232.
- 671 Rushworth, M., Behrens, T., Johansen-Berg, H., 2005. Connection patterns distinguish 3 regions of
672 human parietal cortex. *Cerebral cortex* 16, 1418–1430.
- 673 Schilling, K.G., Nath, V., Hansen, C., Parvathaneni, P., Blaber, J., Gao, Y., Neher, P., Aydogan, D.B.,
674 Shi, Y., Ocampo-Pineda, M., et al., 2018. Limits to anatomical accuracy of diffusion tractography
675 using modern approaches. *bioRxiv* , 392571.
- 676 Song, J.W., Mitchell, P.D., Kolasinski, J., Ellen Grant, P., Galaburda, A.M., Takahashi, E., 2014.
677 Asymmetry of white matter pathways in developing human brains. *Cerebral cortex* 25, 2883–2893.
- 678 Sotiropoulos, S.N., Zalesky, A., 2017. Building connectomes using diffusion mri: Why, how and but.
679 *NMR in Biomedicine* .
- 680 Tournier, J., Calamante, F., Connelly, A., et al., 2012. Mrtrix: diffusion tractography in crossing fiber
681 regions. *International Journal of Imaging Systems and Technology* 22, 53–66.
- 682 Tournier, J.D., Calamante, F., Connelly, A., 2007. Robust determination of the fibre orientation distribu-
683 tion in diffusion mri: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage*
684 35, 1459–1472.
- 685 Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium,
686 W.M.H., et al., 2013. The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79.
- 687 Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction,
688 in: *International Work-Conference on Artificial Neural Networks*, Springer. pp. 758–770.
- 689 Wang, R., Benner, T., Sorensen, A.G., Wedeen, V.J., 2007. Diffusion toolkit: a software package for
690 diffusion imaging data processing and tractography, in: *Proc Intl Soc Mag Reson Med*, Berlin.
- 691 Wasserthal, J., Neher, P., Maier-Hein, K.H., 2018. Tractseg-fast and accurate white matter tract seg-
692 mentation. *arXiv preprint arXiv:1805.07103* .
- 693 Wisse, L.E., Daugherty, A.M., Olsen, R.K., Berron, D., Carr, V.A., Stark, C.E., Amaral, R.S., Amunts,
694 K., Augustinack, J.C., Bender, A.R., et al., 2017. A harmonized segmentation protocol for hippocampal
695 and parahippocampal subregions: Why do we need one and what are the key goals? *Hippocampus*
696 27, 3–11.
- 697 Woolrich, M.W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C.,
698 Jenkinson, M., Smith, S.M., 2009. Bayesian analysis of neuroimaging data in fsl. *Neuroimage* 45,

699 S173–S186.

700 Yeatman, J.D., Dougherty, R.F., Myall, N.J., Wandell, B.A., Feldman, H.M., 2012. Tract profiles of
701 white matter properties: automating fiber-tract quantification. *PloS one* 7, e49790.

702 Yeatman, J.D., Richie-Halford, A., Smith, J.K., Keshavan, A., Rokem, A., 2018. A browser-based tool
703 for visualization and analysis of diffusion mri data. *Nature communications* 9, 940.

704 Zhang, F., Wu, W., Ning, L., McAnulty, G., Waber, D., Gagoski, B., Sarill, K., Hamoda, H.M., Song,
705 Y., Cai, W., et al., 2018. Suprathreshold fiber cluster statistics: Leveraging white matter geometry
706 to enhance tractography statistical analysis. *NeuroImage* .