Magnetic Resonance Imaging 64 (2019) 37–48

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.elsevier.com/locate/mri

Original contribution

Tractography and machine learning: Current state and open challenges

Philippe Poulin[a,*], Daniel Jörgens[b], Pierre-Marc Jodoin[a,1], Maxime Descoteaux[a,1]

[a] Department of Computer Science, Université de Sherbrooke, Sherbrooke, Québec, Canada
[b] Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Keywords:
Diffusion MRI
Tractography
Machine learning
Benchmark

ABSTRACT

Supervised machine learning (ML) algorithms have recently been proposed as an alternative to traditional tractography methods in order to address some of their weaknesses. They can be path-based and local-model-free, and easily incorporate anatomical priors to make contextual and non-local decisions that should help the tracking process. ML-based techniques have thus shown promising reconstructions of larger spatial extent of existing white matter bundles, promising reconstructions of less false positives, and promising robustness to known position and shape biases of current tractography techniques. But as of today, none of these ML-based methods have shown conclusive performances or have been adopted as a de facto solution to tractography. One reason for this might be the lack of well-defined and extensive frameworks to train, evaluate, and compare these methods.

In this paper, we describe several datasets and evaluation tools that contain useful features for ML algorithms, along with the various methods proposed in the recent years. We then discuss the strategies that are used to evaluate and compare those methods, as well as their shortcomings. Finally, we describe the particular needs of ML tractography methods and discuss tangible solutions for future works.

1. Introduction

In the field of diffusion magnetic resonance imaging (dMRI), tractography refers to the process of inferring streamline structures that are locally aligned with the underlying white matter (WM) dMRI measurements [1]. A simple approach to obtain such streamlines is an iterative process in which, starting from a seed point, an estimate of the local tissue orientation is determined and followed for a certain step length before repeating the orientation estimation at the new position. The tracking procedure may be deterministic [2,3] (at each point, the algorithm follows the strongest orientation) or probabilistic [4–6] (at each point, the algorithm samples a direction closely aligned with the strongest orientation). Tracking may also be global as some methods recover streamlines all at once [7–9]. In between the local and global methods is the category of shortest-path methods, including front evolution, simulated diffusion, geodesic, and graph-based approaches [1]. Ultimately, the collection of all trajectories created in that way is called a tractogram.

In traditional methods, the estimate of the local tissue fiber orientation is usually inferred from an explicit and local model which fits the (local) diffusion data. These local models include diffusion tensor models [3,10], multi-tensor models [11,12], and other methods that aim at reconstructing the fiber orientation distribution function (fODF) like constrained spherical deconvolution (CSD) [13,14], to name a few. However, the choice of the best model is by itself difficult [15,16], as it depends on various factors such as data acquisition protocol or targeted WM regions, and therefore has a direct influence on the quality of an obtained tractogram [17]. Moreover, traditional methods based on local orientation alone are prone to make common mistakes, such as missing the full spatial extent of bundles and producing a great amount of false positive connections [16].

Another important factor for the performance of a tractography method are the actual rules that regard the progression of a single step as well as simple global properties of an individual streamline. Traditional methods may define several engineered, or “manually-defined”, high-level rules with the aim of improving the anatomical plausibility of the recovered tractogram. Instances of these are constraints on streamline length (i.e. filtering streamlines that are too long or too short), streamline shape(e.g. filtering streamlines with sharp turns), or progression rules that make streamlines “bounce off” the WM border when they are about to leave the WM mask with a certain angle [18,19]. In the same way as modeling noise and artifacts, and defining the right local model, also the design of these high-level rules has a direct impact on the performance of a tractography method [16,17].

* Corresponding author.
E-mail address: Philippe.Poulin2@Usherbrooke.ca (P. Poulin).
[1] Co-last author.

https://doi.org/10.1016/j.mri.2019.04.013
Received 4 January 2019; Received in revised form 29 April 2019; Accepted 29 April 2019

To address these inherent difficulties, recent proposals suggest that machine learning (ML) algorithms, supervised or unsupervised, may be used to implicitly learn a local, global or contextual fiber orientation model as well as the tracking procedure. Approaches ranging from the application of self-organizing maps (SOM) [20,21], random forests (RF) [19,22], Multilayer Perceptrons (MLP) [23,24], Gated Recurrent Units (GRU) [25–27], as well as Convolutional Neural Networks (CNN) [28] and Autoencoders [29], have been employed at the core of tractography to drive streamline progression. Apart from the differences in their underlying architecture, these ML methods differ substantially in aspects of the exact problem formulation, e.g. definition of the input data to the model, modeling the predictions as a regression [25,29] or classification problem [22,23], or even the general tractography approach, i.e. whole-brain [25,26] or bundle-specific [27–30]. The fact alone that these approaches differ in several aspects, makes it difficult to draw conclusions on the value of each of the individual modeling choices.

Furthermore, while the above mentioned approaches constitute the main ideas for applying ML directly to the process of tractography, machine learning and especially deep learning (DL) methods have been applied in related fields. Stacked U-Nets were proposed to segment the volume of individual white matter bundles from images of fODF peaks [31]. It was also suggested to predict fiber orientations from raw diffusion data based on convolutional neural networks (CNN) [32]. Several ideas for streamline clustering or streamline segmentation have been proposed, including a CNN based on landmark distances [33], a long short-term memory (LSTM)-based siamese network for rotation invariant streamline segmentation [34], and a CNN approach for streamline clustering based on the sequence of their coordinates [35,36]. Even though the mentioned works are closely related to tractography and contribute to the common goal of improved analysis of the white matter anatomy of the human brain, we restrict our focus exclusively on the direct application of ML (and especially DL) for tractography, with the explicit goal of producing streamlines and addressing the weaknesses of traditional methods. For that reason, we refer the interested reader to the respective references for more details.

An important factor for effectively advancing this field of research is a common and appropriate methodology for training and evaluating the performance of different approaches, which is currently lacking. Over the years, multiple challenges have been proposed to assess the performance of conventional tractography methods, and a clear and exhaustive review is provided by [15]. However, we argue that the design of these challenges is typically inappropriate for ML methods. In fact, the *2015 ISMRM Tractography Challenge* [16] (along with the *Tractometer* evaluation tool [17]) has been adopted as the tool of choice for benchmarking new ML tractography pipelines [19,24–26]. Unfortunately, several inherent flaws arising specifically in the context of ML make it difficult to perform a fair comparison between the results obtained from different ML pipelines. In particular, diffusion data preprocessing is left to participants (textbfdissimilar inputs), tracking seeds and a tracking mask are not always given (**varying test environment**), the test diffusion volume is sometimes used for training (**data contamination**), training streamlines are not provided (**disparate training data**), and testing on a single synthetic subject means that any computed estimator of a model's performance is unreliable (**small sample size**). Against the background of a prospectively increasing number of ML-based approaches tackling the problem of tractography, a carefully designed evaluation framework that appropriately addresses the specific requirements of ML methods has the potential to support and facilitate research in this field in the upcoming years.

In this paper, we follow a threefold strategy. First, we introduce the currently available datasets and evaluation tools along with useful features and weaknesses regarding machine learning. Then, we provide a comprehensive review of existing ML-based tractography approaches and derive a set of key concepts distinguishing them from each other. Subsequently, we identify and discuss the strategies for evaluation of tractography pipelines and identify issues and limitations arising when applied to ML-based tractography methods. We finally describe important features for an appropriate evaluation framework the community ought to adopt in the near future to better promote data-driven streamline tractography and point out the potential advantages for research in data-driven streamline tractography.

2. Annotated datasets and evaluation tools

Over the years, many diffusion MRI datasets were produced and annotated, either as part of a challenge or research papers. In this section, we overview several datasets that have been used to train and/or validate supervised learning algorithms for tractography. Specifically, we selected datasets that offer both diffusion data and streamlines. Selected datasets also needed to have either clearly defined evaluation metrics, or to be large enough (more than 50 subjects) to be considered as standalone training sets. We include datasets that are either publicly available or simply mentioned in a research paper without a public release.

We excluded datasets or challenges focused on non-human anatomy (e.g. rat or macaque), where the ground truth is harder to define and results might be harder to generalize to human anatomy (for data-driven algorithms), like the *2018 VOTEM Challenge* [37] (my.vanderbilt.edu/votem/). Moreover, we left out datasets focused only on pathological cases like the *2015 DTI Challenge* [38], because we consider it too early for data-driven tractography algorithms, at least until more conclusive results on healthy subjects. We also excluded tractography atlases when tracking was done on a single diffusion volume, usually averaged over multiple subjects (e.g. HCP842 [39]), because results tend to be overly smooth and unsuited for ML methods. However, we include a recent case when tracking was done for each subject: the 100-subjects WM atlas of [40].

While all the selected datasets are useful in one way or another for data-driven methods, they differ in multiple ways, which are detailed in the following subsections and summarized in Table 1. The listed

Table 1
Annotated datasets.

Name	Year	Public	Real	Human	Subjects	Bundles	GT	Metrics	Split
Fibercup [45]	2009	☑	☑		1	7	☑	☑	
Simulated Fibercup [46]	2012	☑			1	7	☑	☑	
Tracula [47]	2011		☑	☑	67	18	☑		
HARDI 2012 [48]	2012	☑			2	7	☑	☑	☑
HARDI 2013 [49]	2013	☑			2	20	☑	☑	☑
ISMRM 2015 [16,17]	2015	☑		☑	1	25	☑	☑	
HAMLET [30]	2018		☑	☑	83	12			☑
PyT (BIL&GIN) [50]	2018		☑	☑	410	2	☑		
BST (BIL&GIN) [51]	2018		☑	☑	39	5	☑	☑	
TractSeg (HCP) [31]	2018	☑	☑	☑	105	72	☑		
Zhang et al. (HCP) [40]	2018		☑	☑	100	58 + 198	☑		

properties are the following:

- **Name**: The dataset name and reference
- **Year**: The year of publication of the dataset or paper using the dataset
- **Public**: Is the dataset (diffusion data and streamlines) publicly available?
- **Real**: Is the diffusion data a real acquisition or is it simulated?
- **Human**: Does the diffusion data represent the human brain anatomy?
- **Subjects**: The number of subjects or acquisitions
- **Bundles**: The number of bundles or tracks (if streamlines are available)
- **GT**: Is a ground truth known? For real acquisitions, streamlines validated by a human expert (e.g. neuroanatomist) are considered as GT despite the fact that these annotations are subject to inter-rater and intra-rater variations.
- **Metrics**: Well-defined evaluation metrics are available with this dataset.
- **Split**: Is the dataset split into a training and testing set that future works can rely on?

Note that the notion of "ground truth" refers to an indisputable biologically-validated label assigned to an observed variable. In medical imaging, such ground truth may be obtained with a biopsy [41], throughout careful complementary analysis [42] or by having several experts agreeing on a given diagnostic [43]. Unfortunately, such restrictive definition of a ground truth is unreachable most of the time, especially for white matter tracks obtained from tractography, where no expert can truly assess the existence (or non-existence) of a given streamline in a human brain from MRI images only. In fact, only synthetically-generated streamlines or man-made phantoms can be considered as real "ground truth". Despite that, for the purpose of this paper, we also use the term "ground truth" for any data that has been manually validated by a human expert, typically a neuro-anatomist. In the medical imaging field, this annotated data would be called a *gold standard*, while in the artificial intelligence community, it might be called *weakly annotated data*. Although such annotations do not meet the fundamental definition of a ground truth, it is nonetheless widely accepted by the medical imaging AI community [44].

2.1. The FiberCup dataset and the Tractometer tool

2.1.1. Original FiberCup tractography contest (2009)

[45] proposed the *FiberCup Tractography Contest* [45,52] in conjunction with the 2009 MICCAI conference. The goal was to quantitatively compare tractography methods and algorithms using a clear and reproducible methodology. They built a realistic diffusion MR 7-bundle phantom with varying configurations (crossing, kissing, splitting, bending). The organizers acquired diffusion images with b-values of 2000, 4000, and 6000 s/mm^2, and used isotropic resolutions of 3 mm and 6 mm, resulting in 6 different diffusion datasets. Contestants were provided all datasets (but not the ground truth) and were free to apply any preprocessing they wanted on the diffusion images. Evaluation was done by choosing 16 specific voxels, or seed points, in which a unique fiber bundle is expected. Participants were expected to submit a single fiber bundle for each of those seed voxels. Quantitative evaluation was done by comparing the 16pairs of candidate and ground truth fibers using a symmetric Root Mean Square Error (sRMSE).

While the *FiberCup Tractography Contest* makes a good test case for simple configurations, it does not represent a true human anatomy and does not impose a choice of b-value and preprocessing, which can induce significant differences in data-driven methods. Also, it does not provide any training streamlines, and is thus useful only as a validation tool for ML-based methods. Furthermore, the fact that it contains only one subject makes it hard to evaluate the true generalization capability of an ML method trained and tested on that dataset. However, it is the only dataset that provides seed points in order to have a uniform test environment, which is of utmost importance when comparing ML-based algorithms. In the end, it is unclear if for ML-based methods there would be any correlation between a good performance on the FiberCup contest and good performance on human anatomy.

2.1.2. Tractometer evaluation tool (2013)

In 2013, [17] developed the *Tractometer* evaluation tool, to be used alongside the original FiberCup data, with the aim of providing quantitative measures that better reflect brain connectivity studies. Using a Region of Interest (ROI)-based filtering method, a complete tractogram can be evaluated on global connectivity metrics, such as the number of valid and invalid bundles. Furthermore, they propose two seeding masks: a complete mask (mimicking a brain WM mask), and a ROI mask (mimicking GM-WM interfaces). The tractometer was designed to address the fact that "metrics are too local and vulnerable to the seeds given, and, as a result, do not capture the global *connectivity* behavior of the fiber tracking algorithm" [17].

2.1.3. Simulated FiberCup (2014)

In 2014, [53] proposed a simulated version of the FiberCup, allowing new tracking algorithms to be tested using multiple acquisition parameters [53]. The simulated data can be used alongside the *Tractometer* tool designed for the original phantom.

[46] also developed a synthetic version of the FiberCup dataset, but did not publicly release the data [46]. Unfortunately, with regards to ML methods, the simulated FiberCup dataset suffers from the same shortcomings as the original FiberCup dataset as it contains only one non-human subject whose data is not split a priori into a training and testing set.

2.2. Tracula (2011)

[47] published the Tracula method for automated probabilistic reconstruction of 18 major WM pathways. It uses prior information on the anatomy of bundles from a set of training subjects. The training set was built from 34 schizophrenia patients and 33 healthy controls, using a 1.5 T Siemens scanner as part of a multi-site MIND Clinical Imaging Consortium [54]. The diffusion images include 60 gradient directions acquired with a b-value of 700 s/mm^2, along with 10 b = 0 images, with an isotropic resolution of 2 mm. Whole-brain deterministic tracking was performed, followed by expert manual labeling using ROIs for 18 major WM bundles. The dataset also includes a measure of the inter-rater and intra-rater variability for the left and right uncinate.

To our knowledge, this is the earliest apparition of a large-scale human dataset with expert annotation of streamlines. It is also the only dataset that includes a measure of inter-rater and intra-rater variability, which is a desirable feature for ML methods (also discussed later in Section 4.5). Unfortunately, the complete set of diffusion images and streamlines has been incorporated into the method and is not public.

2.3. HARDI reconstruction challenges

2.3.1. HARDI reconstruction challenge (2012)

[49] organized the *2012 HARDI Reconstruction Challenge* [49] at the ISBI 2012 conference. The goal of the challenge was to quantitatively assess the quality of intra-voxel reconstructions by measuring the predicted number of fiber populations and the angular accuracy of the predicted orientations. A training set was released prior to the challenge, and a test set was used to score the algorithms. As such, the 2012 HARDI dataset contains diffusion images but no streamlines.

Participants could request a custom acquisition (only once) by sending a list of sampling coordinates in q-space, and the organizers would then produce a simulated signal for the given parameters. A $16 \times 16 \times 5$ volume was then produced, containing seven different

bundles attempting to recreate realistic 3-D configurations. The metrics proposed by the authors are ill-posed for ML-based methods because of the limited context available and the focus on local performances. Like the *FiberCup*, it would only be useful as a validation tool given the lack of training streamlines, a limited number of bundles (only seven) and a limited number of non-human subjects (only two).

2.3.2. HARDI reconstruction challenge (2013)

The *2013 HARDI Reconstruction Challenge* [48] was organized one year later at the ISBI 2013 conference. For ML-based methods, three improvements are relevant compared to the 2012 challenge: a more realistic simulation of the diffusion signal, a new evaluation system based on connectivity analyses and a larger set of 20 bundles. Indeed, data-driven methods try to learn an implicit representation without imposing a model on the signal, which means that the signal used for training and testing should be as close as possible to that in clinical practice. Furthermore, the main benefit of data-driven methods is the ability to use context in order to make good predictions in a multitude of configurations, which means they have the potential to particularly improve connectivity analyses. Therefore, it would be a better validation tool for ML-based methods than the *2012 HARDI Reconstruction Challenge*. Nonetheless, the dataset suffers from an inherent limitation as it contains only two non-human subjects.

2.4. ISMRM tractography challenge (2015)

This dataset has been designed for a tractography challenge organized in conjunction with the 2015 ISMRM conference [16]. During the challenge, participants were asked to reconstruct streamlines from a synthetic human-like diffusion-weighted MR dataset which was simulated with the aim of replicating a realistic, clinical-like acquisition, including noise and artifacts. The available data consists of a diffusion dataset with 32 b = 1000 s/mm^2 images and one b = 0 image, with 2 mm isotropic resolution, as well as a T1-like image with 1 mm isotropic resolution. Since all data was generated from an expert segmentation of 25 bundles, in theory, a perfect tracking algorithm should only produce exactly these specific bundles. Unfortunately, as for the HARDI and FiberCup datasets, the *2015 ISMRM Tractography Challenge* contains data from a limited number of subjects (only one) and lacks a clear separation between training and testing data. Nonetheless, in combination with the *Tractometer* tool [17], this dataset has often been used to assess ML-based tractography methods. Fig. 1 shows the data generation process for the challenge.

Once a tractogram has been generated using the challenge diffusion data, the tractometer tool uses a “bundle recognition algorithm” [55] to cluster the streamlines into bundles. The generated bundles are then compared to the ground truth, producing groups of “valid bundles” and “invalid bundles”, depending on which regions of the brain the streamlines connect. Streamlines that do not correspond to a ground truth bundle are classified as “No connections” streamlines. The metrics computed by the modified *Tractometer* for the *Tractography Challenge* are as follows:

- **Valid bundles** (VB): The number of correctly reconstructed ground truth bundles.
- **Invalid Bundles** (IB): The number of reconstructed bundles that do not match any ground truth bundles.
- **Valid Connections** (VC): The ratio of streamlines in valid bundles over the total number of produced streamlines.
- **Invalid Connections** (IC): The ratio of streamlines in invalid bundles over the total number of produced streamlines.
- **No Connections** (NC): The ratio of streamlines that are either too short or do not connect two regions of the cortex over the total number of produced streamlines.
- **Bundle Overlap** (OL): The ratio of ground truth voxels traversed by at least one streamline over the total number of ground truth voxels.
- **Bundle Overreach** (OR): The ratio of voxels traversed by at least one streamline that do not belong to a ground truth voxel over the total number of ground truth voxels.
- **F1-score** (F1): The harmonic mean of recall (OL) and precision (1-OR).

The definition of streamline-oriented metrics (VB, IB, VC, IC, NC) and volume-oriented metrics (OL, OR, F1) means that there is no single number that can fully assess the performance of an algorithm. For example, deterministic methods often score higher on streamline-oriented metrics compared to probabilistic methods. As such, a thorough review of all scores must be performed in order to properly compare algorithms, and in many cases, the choice of an algorithm over another may depend on a specific use-case (e.g. bundle reconstruction vs. connectivity analysis).

2.5. HAMLET (2018)

To validate their method, [30] used a dataset of 83 human subjects from two independent cohorts. The first cohort comprises 55 healthy volunteers, all scanned by a Siemens 3T TIM PRISMA MRI scanner. The second cohort has 28 volunteers scanned with a Siemens TIM TRIO. The first cohort was used for training while the second one was used for testing. Subjects in the second cohort were scanned twice for test-retest experiments, some unique characteristic to that dataset. The reference streamlines were obtained by first tracking the whole brain with global tractography, and then by segmenting the streamlines for 12 bundles with a selection algorithm in MNI space. Unfortunately, the recovered streamlines have not been manually validated by an expert.

2.6. Datasets based on the BIL&GIN database

2.6.1. Bundle-specific tractography (2018)

[51] proposed a bundle-specific tracking method based on anatomical priors that improves tracking in the centrum semiovale crossing regions [51]. Using multiple tractography algorithms, they tracked and segmented five bundles (Arcuate Fasciculus - AF left/right, Corpus Callosum - CC, Pyramidal Tracts - PyT left/right) in 39 subjects from the BIL&GIN database [56]. To compare algorithms, they used an automatic bundle segmentation method based on clear anatomical definitions. In addition, they defined several performance metrics, such as *bundle volume*, *ratio of valid streamlines*, and *efficiency*. However, the tractograms and automatic bundle segmentation procedure were neither made public nor validated by an expert. Such a dataset, along with the evaluation procedure, could be extremely useful to assess if data-driven methods can reliably learn the structure of a specific bundle and reconstruct it in unseen subjects.

2.6.2. A population-based atlas of the human pyramidal tract (2018)

[50] created a streamline dataset of the left and right PyT based on a population of 410 subjects [50], also from the BIL&GIN database [56]. To do so, they combined manual ROIs along the bundles' pathway and the bundle-specific tractography algorithm of [51]. The quality of the segmentations and the high number of subjects would make this a noteworthy training dataset for data-driven methods. Unfortunately for ML methods, only two bundles were examined. Furthermore, while the probability maps of the atlas have been rendered public, the tractograms are still unavailable.

2.7. Datasets based on the HCP database

2.7.1. TractSeg (2018)

[31] proposed a data-driven method for fast WM tract segmentation without tractography [31]. In doing so, they built an impressive dataset of 72 manually-validated bundles for 105 subjects from the Human Connectome Project (HCP) diffusion database [57,58]. Tractograms

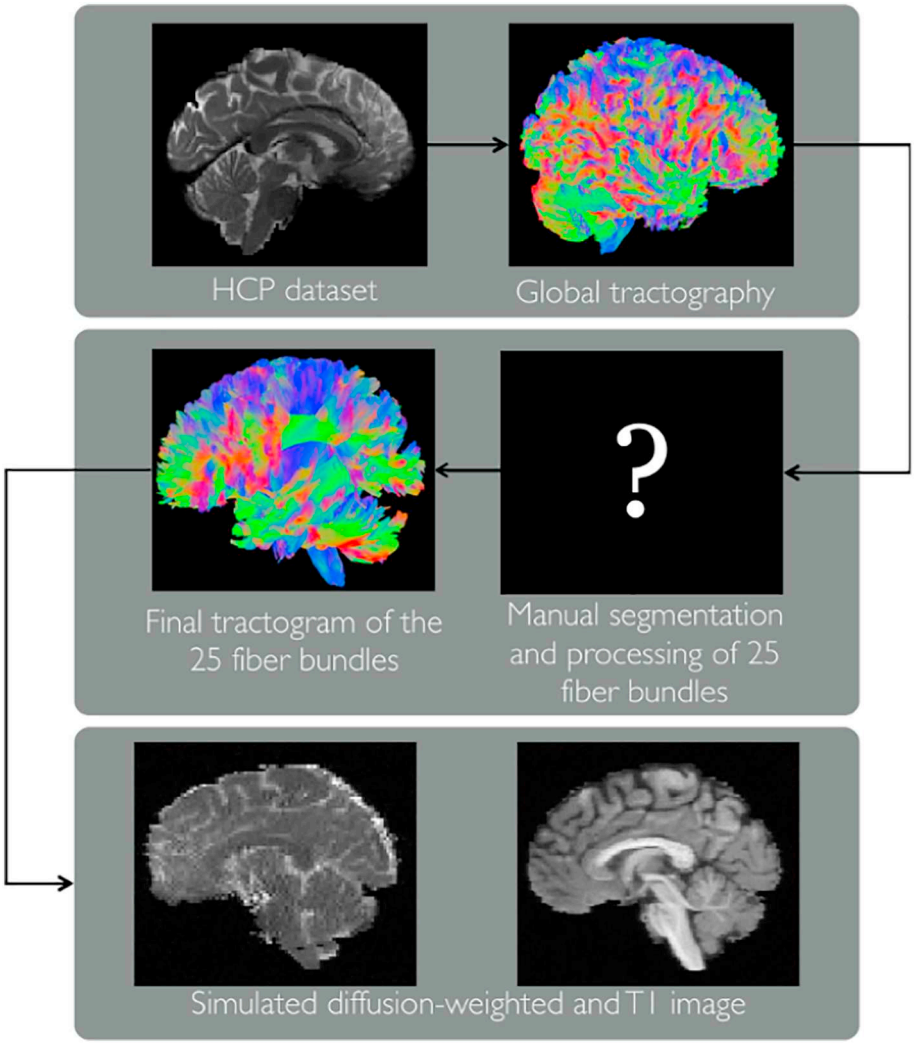

Fig. 1. 2015 ISMRM Tractography Challenge data generation process (Taken from www.tractometer.org).

were obtained via a four-step semi-automatic approach:

1. Tractography (Multi-Shell Multi-Tissue CSD [5])
2. Initial tract extraction (TractQuerier [59])
3. Tract refinement (Manual ROIs [60] + QuickBundles [61])
4. Manual quality control and cleanup

To the best of our knowledge, this is the largest public database to include both diffusion data and reference streamlines. No further pre-processing of the diffusion data is needed because of the standard procedure of [58]. The authors defined volume-oriented metrics such as Dice score [62], but did not offer any streamline-oriented metrics as their method predicts a volume segmentation. The high number of subjects and bundles makes this a remarkable training set.

In a subsequent paper, the same authors re-used a subset of 20 bundles of the TractSeg dataset to train and validate their TOM ML algorithm [28]. However, as for original 72-bundle dataset, the TOM dataset does not come with a predefined set of training and testing data and no formal evaluation protocol that users could rely on has been proposed.

2.7.2. Zhang et al. (2018)

[40] built a WM fiber atlas using 100 HCP subjects. They first generated streamlines for all subjects using a two-tensor unscented Kalman filter method [63], and sampled 10,000 streamlines from each subject after a tractography registration step. Then, using a hierarchical clustering method, the authors generated an initial WM fiber atlas of 800 clusters. Finally, an expert neuroanatomist reviewed the annotations in order to accept or reject each cluster, and provided the correct annotations when the initial annotation was rejected. The final, proposed atlas is comprised of 58 bundles (each composed of multiple clusters), along with "198 short and medium range superficial fiber clusters organized into 16 categories according to the brain lobes they connect" [40].

While the atlas is public, the sampled streamlines from the 100 subjects are all merged into the single template. In order for ML methods to benefit from this dataset, the streamlines would need to be separated back into the space of the particular original subjects. For this reason, we do not consider this dataset to be "public", in the context of machine learning.

3. Machine learning methods for tractography

For this review, we regard all supervised machine learning methods published in peer-reviewed journals, conferences or on arXiv (arxiv.org) and biorXiv (biorxiv.org). We added the requirement that methods needed to be specifically designed for tractography, i.e. with the purpose of predicting a contextual streamline direction (and not reconstructing a local, non-conditional fODF or clustering streamlines). This criterion includes whole-brain as well as bundle-specific

Table 2
Main properties of data-driven methods for tractography.

Method	Model	Temporal context	Spatial context	dMRI input	Prediction	Implicit stop
Neher et al. [19]	RF	1 last direction	50 samples	Resampled DWI	Classification	☑
Poulin et al. [25]	GRU	Full	1 × 1 × 1 voxel	SH	Regression	
Poulin et al. [27]	GRU	Full	1 × 1 × 1 voxel	SH	Regression	
Benou et al. [26]	GRU	Full	1 × 1 × 1 voxel	Resampled DWI	Classification	☑
Jörgens et al. [23]	MLP	2 last directions	1 × 1 × 1 voxel	Raw DWI	Regression	
Wegmayr et al. [24]	MLP	4 last directions	3 × 3 × 3 voxels	SH	Regression	
Wasserthal et al. [28]	CNN	N/A	Entire WM	fODF peaks	Regression	
Reisert et al. [30]	CNN-like	N/A	Entire WM	SH	Regression	☑

RF: Random Forest; MLP: Multilayer perceptron; GRU: Gated recurrent unit; CNN: Convolutional neural network; SH: Spherical harmonics coefficients; fODF: fiber Orientation Distribution Function; Implicit stop: indicates if a method learns its tracking stopping criterion or if it relies on a usual explicit criterion.

tractography methods. A summary of the main properties for all reviewed methods is provided in Table 2.

3.1.1. Random forest classifier

To the best of our knowledge, [22] were the first to propose a machine learning algorithm for (deterministic) tractography [22]. They employ a RF classifier to learn a mapping from raw diffusion measurements to a directional proposal for streamline continuation. After collecting several of such proposals in a local neighborhood of the current streamline position (radius: 25% of the smallest side length of a voxel), these are aggregated in a voting scheme to finally arrive at a single direction in which to grow the streamline.

To define reference streamlines for their experiments, the authors employ several tractography pipelines and train their classifier on each of the resulting tractograms. They determine the best trained model by evaluating the performance of each on a replication of the FiberCup phantom (based on the *Tractometer* metrics of [17]). Finally, comparing the performance of the latter to all other reference pipelines, they report a superior performance of their tracking model over all other approaches. While tractograms were scored on a simulated phantom (i.e. no real anatomy), extended experiments presented in a subsequent paper [19] confirm the superiority of their approach on the *2015 ISMRM Tractography Challenge* dataset (simulated data of a human anatomy).

3.1.2. Gated recurrent unit (GRU) tracking

Hypothesizing "that there are high-order dependencies between" the local orientation at a point of a streamline and the orientations at all other points on the same streamline, [25] proposed a recurrent neural network (RNN) based on a GRU [25] to learn the tracking process. Their method implements an implicit model mapping diffusion measurements to local streamline orientations which not only depends on measurements in a local context, but on all data previously seen along the extent of a particular streamline. As opposed to [19,22], the RNN model is implemented as a regression approach. In their experiments, the authors show that a recurrent model (when trained on reference streamlines obtained using deterministic CSD-based tractography [6]) was able to outperform most of the original submissions in the *2015 ISMRM Tractography Challenge* with respect to the *Tractometer* scores (discussed in Section 2.4).

3.1.3. DeepTracker

In a subsequent paper, [27] again suggested using a GRU, but in a bundle-specific fashion. While the model architecture is very similar, it was trained on a dataset of 37 real subjects, each with a curated set of streamlines for bundles. After training a single model for each of the selected bundles, the authors showed promising results compared to existing methods, perhaps indicative that the difficult task of learning to track streamlines necessitates more data than previously thought.

3.1.4. DeepTract

More recently, [26] proposed a GRU-based recurrent neural network similar to that of [25]. In their method, they directly use the resampled diffusion signal as input to the model (like [19]), in order to estimate a discrete, streamline-specific fODF representation which they refer to as "conditional fODF" (CfODF). Instead of predicting a 3D orientation vector using a regression approach, the authors implement their model as a classifier enabling them to interpret the probabilities obtained for discrete sampled directions (i.e. the classes) as the mentioned CfODF. This fODF-basedformulation further allows for an inherently defined criterion for streamline termination based on the entropy of the CfODF. The proposed model can be employed for both deterministic and probabilistic tractography.

Like [25], the authors trained and tested their method on the *2015 ISMRM Tractography Challenge* dataset. They report results after training their method on the dataset ground truth as well as on streamlines obtained with the MITK diffusion tool [64].

3.1.5. Multi-layer perceptron point-wise prediction

[23] propose a multi-layer perceptron (MLP) to predict the next step of a streamline. Like [19,22,25], their method takes as input the diffusion signal and thus avoids explicit dMRI model-fitting. The authors implemented different configurations of their proposed MLP such as three different input scenarios (point-wise input vs region-wise input with and without considering previous orientations), different approaches to aggregate the output (maximum likelihood, mathematical expectation of the categorical prediction and regression) as well as the voting scheme proposed by [22]. Results reveal that the best configurations are those having the previous two directions included in the input of the network thus showing that temporal context is a key component for data-driven tractography. Also, the regression and classification approaches led to similar results and the use of region-wise information did not provide any substantial improvement over the use of point-wise information.

Like [25,26], the authors trained and tested their method on the *2015 ISMRM Tractography Challenge* dataset (but did not use the *Tractometer* tool). Unfortunately, they did not estimate the tracking capabilities of their method as they only measured point-wise angular errors when predicting the next step of a streamline.

3.1.6. Multi-layer perceptron regression tracking

A similar approach suggested by [24] employs a MLP to predict the next direction of a streamline through regression. At each point, the input of the model is given by all diffusion measurements in a cubic neighborhood, along with a certain number of previous steps for the current streamline. In that way, the authors provide the ML model directly with diffusion information in a local neighborhood (spatial context) as well as a notion of "history" of the current streamline (temporal context). Defining their reference streamlines as tractograms obtained with a standard tractography method from in vivo datasets, they train their model on three subjects from the HCP database.

Experimental validations on the *2015 ISMRM Tractography Challenge* dataset reveal that their model outperforms some ML methods [19,25] in most *Tractometer* metrics. However, as demonstrated by low overlap scores, the authors acknowledge that their model produces "rather confined bundles with little spread", especially in contrast to [19,25]. While the strength of this model is to explicitly provide information from a local neighborhood, like for [23], the notion of context along the streamline is limited and needs to be defined before training. Since the ideal temporalcontext (in terms of streamline length, or steps) is still unknown, this could potentially prohibit the model from taking advantage of all information relevant to streamline continuation.

3.1.7. Tract orientation mapping using an encoder-decoder CNN

[28] proposed a data-driven, bundle-specific tracking method. As opposed to the other ML methods reported in this paper, the authors do not try to directly reconstruct streamlines per se. Instead, their proposed *Tract Orientation Mapping* (TOM) method predicts bundle-specific fODF peaks that are then used by a deterministic tracking method. First, CSD is used to extract three principal directions in all WM voxels. Then, a U-Net CNN [65] is trained to map these fODF peaks to bundle-specific peaks, i.e. peaks that are only relevant for the streamlines of a given bundle. Their CNN takes as input 9 channels (the three fODF peaks) and outputs 60 channels, i.e. a 3D bundle-specific fODF vector for each of the 20 bundles they are looking to recover. While the recovered bundle-specificpeaks can be used in different ways, the authors show that using them directly as input to a deterministic MITK diffusion tractography gives some of the best results. The approach was trained and tested on 105 HCP subjects, each with reference streamlines produced by a semi-automatic dissection of 20 large WM bundles (which they recently rendered public [31]).

3.1.8. HAMLET

In a similar line of thought, in their HAMLET project (*Hierarchical Harmonic Filters for Learning Tracts from Diffusion MRI*) [30] map raw spherical harmonics of order 2 to a spherical tensor field. In that sense, like etwasserthal2018tract, their ML method does not output streamlines but instead voxel-wise bundle-specific tensors that can subsequently be used as input to a classical tractography method. The magnitude of the produced 3×3 tensor indicates thepresence of a specific bundle whereas the tensor orientation predicts the local streamline direction. Their method implements a multi-resolution CNN with rotation covariant convolution operations. They trained and tested their method on two in-house datasets comprising a total of 83 human subjects. The 12 bundles and their associated reference streamlines have been obtained with global tractography and automatic bundle selection method. Unfortunately, the reference data was not manually validated by a human expert, and they did not perform any comparisons against other tractography methods.

4. Results & discussion

4.1. Results on the 2015 ISMRM tractography challenge

The *2015 ISMRM Tractography challenge* is the only dataset that has been used to assess performance of several data-driven tractography methods and is thus, as of today, the only available common ground on which to compare methods. It was used by four different papers namely, the Random-Forest of [19], the GRU of [25,26], and the MLP of [24]. Experimental results reported by the authors have been transcribed in Table 3, and compared with original submissions in Fig. 2. Note that the metrics marked as *not available* (N/A) are those the authors did not report in their original paper.

As can be seen, results vary a lot and there is no clear trend showing which method performs best, especially given the nature of the evaluation metrics. As mentioned in Section 2.4, methods can be evaluated using both *streamline-oriented* metrics and *volume-oriented* metrics, which are not always correlated. For example, a method may have a large number of valid connections but a low overlap (like the MLP of [24]) which means that although the model was able to recover most valid bundles, the generated streamlines do not properly cover the spatial extent of those bundles. Also, a method can be more conservative and score best in terms of invalid connections and overreach like the GRU of [26], but at the same time have a low ratio of valid connections and a poor bundle overlap. On the other hand, the Random-Forest of [19] does not score best in any category, but is competitive according to all metrics (its large F1-score underlines that it is a more balanced method compared to MLP and DeepTract). On top of that, all methods were trained and evaluated differently, so any comparison based on the reported results should be done with extreme care.

4.2. The 2015 ISMRM tractography challenge as an evaluation tool for ML algorithms

As mentioned before, the *2015 ISMRM Tractography Challenge* has been adopted as the de facto evaluation tool to compare ML tractography methods. However, the strengths and weaknesses of that tool should be thoroughly reviewed to understand and trust any technique reporting results with it. In this section, we present what we consider to be important issues with the way in which this tool has been used to assess the performance of data-driven methods. In particular, we detail the discrepancies between the four ML-based methods, differences that may explain some of the results in Table 3 and potentially undermine any conclusion that one could draw from it. Let us mention that some of these issues with the 2015 ISMRM dataset are typical for the field of tractography as a whole.

Table 4 presents a summary of the differences in how the tool is used. Note that the *not available* (N/A) mark is used for any information the authors did not mention in their original paper.

4.2.1. Dissimilar inputs

The four ML methods use a different preprocessing pipeline. Among the proposed algorithms, two applied MRtrix's *dwidenoise* or *dwipreproc* (www.mrtrix.org), another one denoised using [66] and corrected for eddy currents and head motion, and another one did not apply any preprocessing at all. Moreover, some used the diffusion signal directly as input, while others resampled it to a specific number of gradient directions. In some cases, spherical harmonics were fitted to the signal and the SH coefficients were fed as input to the model. Finally, the non-recurrent models are also given a variable number of previous streamline directions as input.

The output of each of these pipelines contain various degrees of information. For example, fODF peaks are in theory already aligned with the major WM pathways, and information may be lost depending on the specific model used to recover the peaks from the diffusion signal. On the other hand, using the raw diffusion signal might contain more information but is more difficult to understand and process, and thus a data-driven model might require more capacity to use such an input. Without a thorough investigation of the information contained in each output, any variations in the *Tractometer* results could be attributed to the variations in preprocessing. Since we currently do not have any indication of what is useful for data-driven algorithms, it is impossible to compare ML methods if they do not use the same input data.

4.2.2. Varying test environment

Since no white matter mask is provided, it must be computed by each participant in case it is needed for tracking. Out of the four ML methods that were evaluated on the challenge, two needed WM masks; one used the ground truth mask, and the other did not mention how the mask was computed. Furthermore, since no tracking seeds are supplied with the data either, their arrangement entirely depends on the WM mask (and the number of seeds per voxel, which is also not given).

Given the nature of streamline tractography, small variations of the

Table 3
Tractometer results. The Bundles and Connections (%) metrics are *streamline-oriented metrics* whereas the Avg. bundle (%) metrics are *volume-oriented metrics*.

Model	Bundles		Connections (%)			Avg. bundle (%)		
	Valid	Invalid	Valid	Invalid	No connection	Overlap	Overreach	F1-score
Random-forest [19]	**23**	94	52	N/A	N/A	59	37	61
GRU [25]	**23**	130	42	46	**13**	**64**	35	**65**
MLP [24]	**23**	57	**72**	N/A	N/A	16	28	26
GRU (DeepTract) [26]	**23**	**51**	41	**33**	23	34	**17**	44

tracking mask or the tracking seeds could have a substantial impact on the resulting streamlines and by that also on the obtained evaluation metrics. Also, even though computing a stopping criterion within the algorithm is a worthy improvement, it is a different task than tracking, and should be evaluated separately. Consequently, all methods should be provided the same tracking mask and seeds to reduce as much as possible the number of free variables during evaluation.

4.2.3. Data contamination

The use of ML methods requires special care when dealing with available data. Since machine learning models are obtained by deriving implicit rules **directly from given data** (i.e. *training data*), testing the true generalization capabilities of these rules must be done using a **different and unseen set of data** (i.e. *test data*).

Two methods suffer from data contamination, or *leakage* [67]: the GRU in [25] and the DeepTract model in [26]. Here, data contamination refers to the usage of the same diffusion data for training and testing. This means that the true generalization capabilities of the tested method on new, thus unseen subjects are still unknown, since the model has already seen the specific diffusion patterns that are needed in order to "explore" at test time, and therefore has been given an "unfair" advantage.

4.2.4. Disparate training data

All methods used different reference streamlines and subjects for training. As mentioned earlier, some employed the test diffusion data directly, while others relied on a varying number of subjects from the HCP database. Two methods used deterministic CSD tracking [6] to generate reference streamlines, one used QBI tracking [68] (probabilistic) and the last one used iFOD tracking [5] (also probabilistic). In order to provide a uniform basis for comparison, the same comprehensive streamline training set should be available to every algorithm.

4.2.5. Simulation as a substitute for human acquisition

While the diffusion signal of the 2015 ISMRM dataset is typical of that of a human brain, it is nonetheless obtained through simulation. As such, results on that dataset should not be seen as a measure of future performance on real human subjects, at least not without further empirical evaluation. Furthermore, at the given resolution and using this particular configuration of 25 bundles, false positive streamlines that would otherwise be plausible given the underlying anatomy of a real scan might be impossible to avoid. Indeed, some authors tried training their models using the ground truth bundles, and still produced over 50 invalid bundles in both cases [19,26].

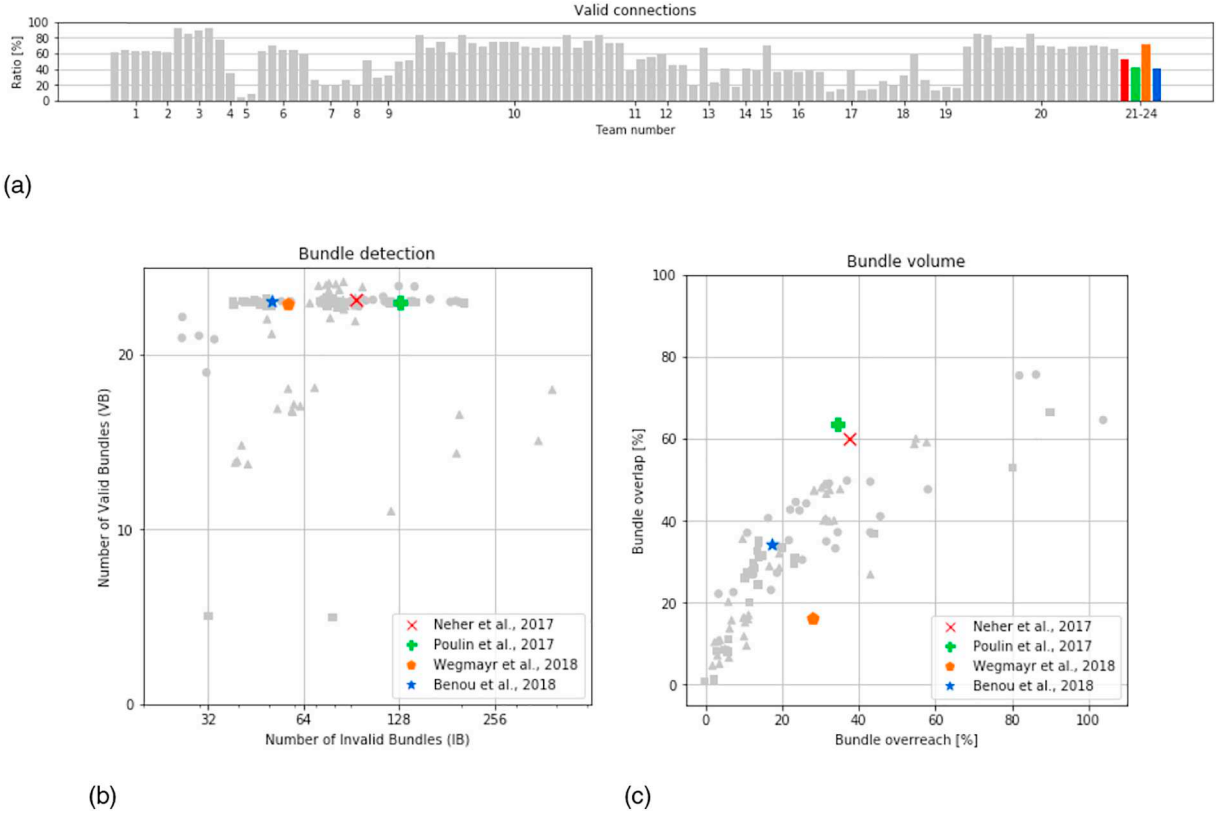

Fig. 2. *2015 ISMRM Tractography Challenge* original submissions (1 – 20) and new results (21–24).

Table 4
Differences in data.

Method	Preprocessing	WM mask	Training subjects	Reference streamlines
Random-Forest [19]	*dwidenoise* + *dwipreproc*	Not needed	5 HCP subjects	CSD (deterministic)
GRU [25]	None	Ground Truth	Challenge subject	CSD (deterministic)
MLP [24]	*dwipreproc*	N/A	3 HCP subjects	iFOD (probabilistic)
DeepTract [26]	N/A	Not needed	Challenge subject	Q-Ball (probabilistic)

4.2.6. Small sample size

The *2015 ISMRM Tractography Challenge* dataset has only one subject, which makes it hard to assess the future performance of a data-driven algorithm [69]. In order to compute unbiased estimates of future performance, a richer test set with more subjects is needed. Also, given more subjects, bootstrapping methods [70] (i.e. sampling with replacement) could help to build more accurate estimators.

4.3. Other results

Some authors report local performance measures, such as the mean angular error [23]. However, local metrics do not take into account compounding errors, which can have a major effect on global structure. Consequently, global evaluation metrics should be preferred.

Tractography papers often report a visual evaluation on unseen, in vivo subjects, as a qualitative evaluation. For example, Figs. 3 and 4 compare some of the proposed data-driven approaches with standard tractography methods on white matter bundles with known anatomy. However, in absence of a ground truth or the expertise of a neuroanatomist, it is hard to draw definitive conclusions on the quality of such results. In addition, [30] presented correlation plots to assess reproducibility, but only offered qualitative comparisons with the reference streamlines without any quantitative results. To gain trust in these data-driven methods, a more rigorous approach is needed.

Finally, most ML methods offer a reduction in computation time compared to traditional methods. This is a non-negligible benefit, should these methods be adopted in practice.

4.4. Limitations of machine learning algorithms for tractography

4.4.1. Bundle exploration

Machine learning algorithms, by definition, learn to reproduce structured information from a given training dataset. In the case of tractography, the structure of bundles is an integral part of the training dataset, and bundle reconstruction is indeed at the core of all the proposed data-driven tractography methods. As such, trained methods should be expected to be able to reproduce those bundles. However, they should not be expected to be able to "extrapolate" the brain and "discover" new bundles unseen in training. In fact, it is still unknown if data-driven methods can extract enough information from the diffusion signal to solve all possible fiber configurations, even when provided with temporal or spatial context.

Until it is shown otherwise, machine learning algorithms for tractography should not be expected to reconstruct bundles unseen during training, and thus should not be trusted to extend their reconstruction capabilities to new connections or new brain regions like traditional methods.

4.4.2. Signal noise

Working with diffusion MRI can be difficult because it usually provides a low SNR compared to other modalities, and can be subject to different noise distributions, such as non-central Chi distributions (which includes the Rician distribution), or the Gaussian distribution [71–73]. Having access to lots of training data from different sources can mitigate this problem, but this is still a challenge for data-driven methods, as it requires large training sets covering the full spectrum of

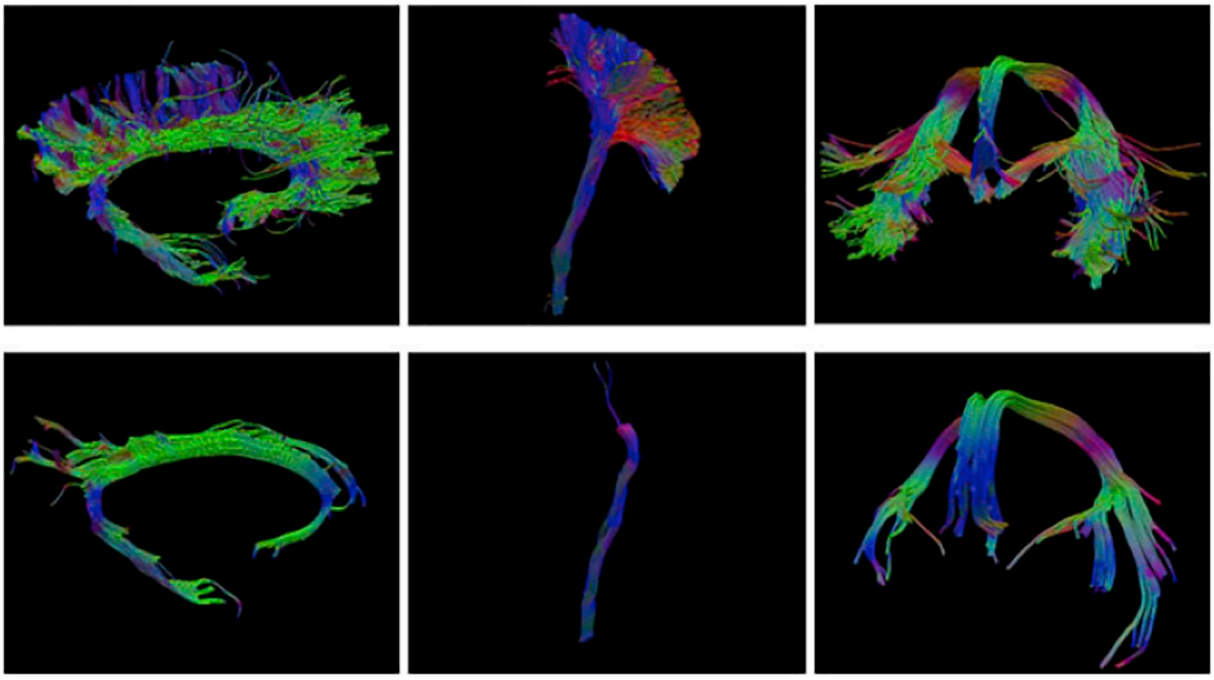

Fig. 3. Comparison between the RF of Neher et al. (top row), and classical deterministic CSD streamline tractography (bottom row). Results obtained on HCP subject 992,774. (*Taken from* [19] *with authorization from the authors*).

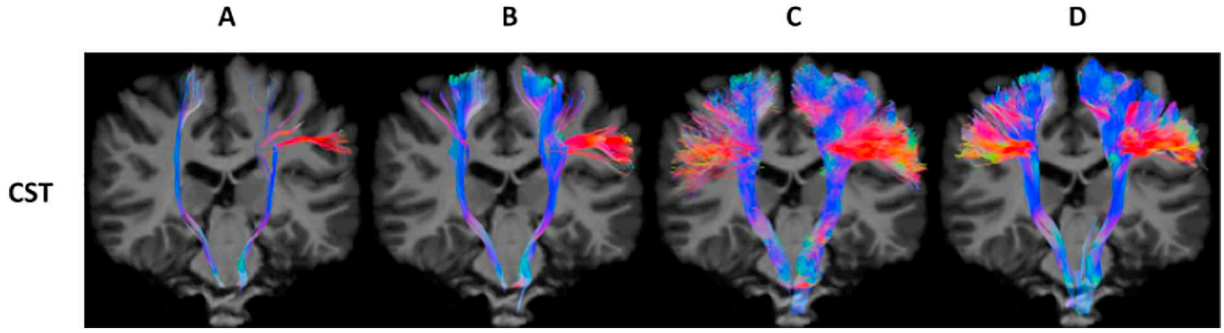

Fig. 4. Comparison of various tracking methods: A: Deterministic, B: Deterministic Bundle-Specific (DET-BST) [51], C: Probabilistic particle filter BST (PROB-PF-BST) [18], D: DeepTracker [27]. Results obtained on a BIL&GIN subject. (*Taken from* [27] *with authorization from the authors*).

potential noise profiles. Standard diffusion MRI pre-processing software is especially useful in this case to remove as much noise as possible and make the data more manageable for data-driven models. However, if data seen at test time is subject to a different noise distribution than what was seen while training, the model might fail where a classic tractography method would have worked.

4.4.3. Predetermined input dimensionality

Varying parameters in diffusion MRI protocols, such as the number of gradient directions, their orientation, and the b-value (or multiple b-values) used to probe the tissue structures, can affect data-driven models that directly rely on the raw data as input. Given that ML models generally take a vector of information with fixed dimensionality as input, the number of gradient directions would predetermined and could not be changed after training of such models. Changing the orientation of the gradients or changing the b-value after training would also change the input signal distribution, and is therefore generally not supported by machine learning models. That is to say, the scanning protocol should stay the same between training and testing.

Introducing an intermediate step of processing the raw diffusion signal, e.g. re-sampling the diffusion signal [19,26] or fitting spherical harmonics [24,25,27,30], is on way to overcome the fixed number of gradient directions and their orientations, but it is still unknown how the model will be affected by changing the underlying measured signal. Furthermore, supporting different b-values (or even a varying number of shells) remains an open problem for models using a fixed-size input.

4.5. Proposed guidelines for a data-driven tractography evaluation framework

Considering the ML tractography evaluation issues previously underlined, we discuss in this section the fundamental elements of a better framework we believe the community should adopt in the upcoming years. We start with the essential characteristics such a framework should have, followed with useful features.

4.5.1. Essential characteristics

First and foremost, an ideal data-driven tractography evaluation framework should come with a public and free-to-use dataset that anyone could easily rely on. The dataset should include images of real human acquisitions along with a careful expert selection of ground truth streamlines. It is important to avoid any bias towards a specific tractography algorithm. In order to achieve this, the streamlines could be first generated by a large number of different (and ideally orthogonal) deterministic, probabilistic and global algorithms and then segmented by expert annotators according to strict anatomical definitions for a given number of bundles. While such manual annotation would be tedious, time consuming and even error prone, we consider this an indispensable step towards building a realistic and useful dataset for ML-based development. The need for such a gold standard that quantifies human variability is well-known in other fields, such as automatic image segmentation, cell counting or in machine learning [74–77]. Despite the fact that simulated brain images come with a pixel-accurate set of ground truth streamlines that can be generated in a matter of seconds, by definition synthetic diffusion signals are over-simplistic pictures of real data and, as such, cannot provide any guarantee of subsequent performance for data-driven methods on real data.

Although there is no consensus regarding the most desirable features a ML tractography algorithm should have and how it should be evaluated, by its very nature, any ML evaluation framework should aim at measuring how an algorithm can faithfully reproduce a task it was trained for. As such, a reasonable dataset should include a sufficiently large number of well-separated training and testing images. Thus, statistics resulting from such a dataset would not suffer from contamination and the reported metrics would be reliable and unbiased estimates of the true generalization power of a ML algorithm. In addition, to ensure that the observed differences between multiple algorithms are resulting from the intrinsic properties of the model and not caused by some feature of the evaluation framework, the number of free variables should be reduced to a minimum. Consequently, the tracking masks and seeds should be provided together with clearly preprocessed diffusion data, so that the proposed methods can be evaluated in equal conditions. There should be multiple “classes” of input data, depending on whether an algorithm supports DWI samples, SH coefficients or fODF peaks. Furthermore, the initial diffusion signal should have the same statistical properties for the training and the testing set. Finally, the acquired images should ideally be acquired at different MRI scanners with different acquisition protocols in order to avoid overfitting issues.

Evaluation metrics should also be bound to the purpose of tractography algorithms. Considering that tractography is mostly used for bundle reconstruction, tractometry studies and connectivity analyses, an ideal evaluation framework should include two sets of metrics: 1) metrics measuring how a ML method can faithfully reproduce a set of predefined bundles it was trained to recover (tractometry), and 2) metrics measuring how it can connect matching regions of the brain, i.e. produce valid connections (connectivity). Furthermore, since many applications use tractography algorithms to produce a large number of streamlines (with many false positives), which are then filtered out by a post-processing algorithm such as RecoBundles [55], the framework should report results before and after post-processing. This would underline the true recall power of a data-driven algorithm, which is a fundamental characteristic of tract-based and connectivity-based applications [16].

Lastly, the size of an ideal dataset is of primary importance. While a small-sized dataset could be prone to overfitting, it would be costly to create a very large dataset and also difficult to ensure a coherent manual annotation. One rule of thumb that can be used to identify the “correct” size of a dataset is through the inspection of the learning curve of several ML models [78]. These curves show the model performance as a function of the training sample size. Typically, the

performance of several models saturates for a sufficient dataset size. Although imperfect, this procedure is a good heuristic for estimating the size of the dataset.

4.5.2. Other useful features

Despite any thorough manual annotation protocol, manually annotated bundles can be subject to non-negligible inter-rater and intra-rater variability. As such, a useful characteristic of a ML tractography dataset would be a measure of those variations. This would be obtained by having several experts annotating the dataset, and at least one expert annotating it twice or more times. Such measures would provide a minimal bound beyond which a data-driven algorithm could be considered "as good as an expert". Another very useful tool would be an openly accessible online evaluation system. Given such a system, people could upload their test results in order to compare them with the test ground truth. In that way, an automatic ranking procedure similar to that of Kaggle could be used to sort various ML algorithms based on their achieved scores. While no ranking method is perfect, it would nonetheless provide a common evaluation framework that people could rely on.

An ideal dataset would also cover the whole field of diffusion MRI acquisition protocols, from HCP-like research acquisitions to clinical acquisitions. It would include single b-value as well as multiple b-values data, along with more sophisticated acquisition protocols such as b-tensor encoding. It would also need low resolution images together with high-resolution images. Since data harmonization is also a problem for data-driven algorithms, acquisition from several sites are needed for test-retest studies. Annotated pathological cases would complete the dataset by allowing careful preliminary studies on how ML-based methods can be relied on in unhealthy patients.

Finally, since tractography is used more and more in pre-clinical applications, a subset of manually annotated rodent or macaque brains would be of great interest to train and test future ML algorithms (like the *2018 VOTEM Challenge* [37], for example).

This is, of course, the ultimate wish list. But, in the era of open data and open science, it needs to be done by the community, for the community. We can already see this work in progress with more and more accessible and reproducible data being published every year.

5. Conclusion

In this paper, we provided an exhaustive review of the current state of the art of machine learning methods in the field of tractography. We described the existing datasets that comprise both diffusion data and reference streamlines, which could generally be useful for new tracking methods based on ML. In particular, we thoroughly examined the widely used evaluation tool for data-driven tracking methods, the *2015 ISMRM Tractography Challenge*, and detailed flaws and shortcoming when used to assess data-driven algorithms. Based on our findings, we suggested good practices that we believe would foster the development of a new evaluation framework for ML-based tractography methods with the potential to effectively advance this field of research.

There is no doubt that machine learning tractography will have an important role to play in the future to solve some of the open problems of tractography. At the moment, however, all existing methods show theoretical potential and in limited test cases. Methods have yet to make solid demonstrations of their performance and efficiency in practice. There is still no ML-based tractography tool that is a scalable and usable on any given diffusion MRI dataset. This is true for healthy datasets but even more so for pathological brains. Hence, it is fair to say that ML-based tractography is still at its infancy and is not ready for "prime-time", but is nonetheless a very fertile field of research to make meaningful contributions to the field of connectivity mapping.

Funding information

FRQNT; Université de Sherbrooke Institutional Chair in Neuroinformatics; NSERC Discovery grant from Pr Descoteaux and Jodoin.

References

[1] Jeurissen B, Descoteaux M, Mori S, Leemans A. Diffusion MRI fiber tractography of the brain. NMR Biomed 2017:e3785.
[2] Yeh FC, Verstynen TD, Wang Y, Fernández-Miranda JC, Tseng WYI. Deterministic diffusion fiber tracking improved by quantitative anisotropy. PloS one 2013;8(11):e80713.
[3] Basser PJ, Pajevic S, Pierpaoli C, Duda J, Aldroubi A. In vivo fiber tractography using DT-MRI data. Magn Reson Med 2000;44(4):625–32.
[4] Behrens TE, Berg HJ, Jbabdi S, Rushworth MF, Woolrich MW. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? Neuroimage 2007;34(1):144–55.
[5] Tournier JD, Calamante F, Connelly A. Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. Proceedings of the international society for magnetic resonance in medicine. vol. 18. 2010. p. 1670.
[6] Tournier JD, Calamante F, Connelly A. MRtrix: diffusion tractography in crossing fiber regions. Int J Imaging Syst Technol 2012;22(1):53–66.
[7] Reisert M, Mader I, Anastasopoulos C, Weigel M, Schnell S, Kiselev V. Global fiber reconstruction becomes practical. Neuroimage 2011;54(2):955–62.
[8] Mangin JF, Fillard P, Cointepas Y, Le Bihan D, Frouin V, Poupon C. Toward global tractography. Neuroimage 2013;80:290–6.
[9] Jbabdi S, Woolrich MW, Andersson JL, Behrens T. A Bayesian framework for global tractography. Neuroimage 2007;37(1):116–29.
[10] Pierpaoli C, Jezzard P, Basser PJ, Barnett A, Di Chiro G. Diffusion tensor MR imaging of the human brain. Radiology 1996;201(3):637–48.
[11] Caan MW, Khedoe HG, Poot DH, Arjan J, Olabarriaga SD, Grimbergen KA, et al. Estimation of diffusion properties in crossing fiber bundles. IEEE Trans Med Imaging 2010;29(8):1504–15.
[12] Malcolm JG, Shenton ME, Rathi Y. Filtered multitensor tractography. IEEE Trans Med Imaging 2010;29(9):1664–75.
[13] Tournier JD, Calamante F, Gadian DG, Connelly A. Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. NeuroImage 2004;23(3):1176–85.
[14] Descoteaux M, Deriche R, Knösche TR, Anwander A. Deterministic and probabilistic tractography based on complex fibre orientation distributions. IEEE Trans Med Imaging 2009 feb;28(2):269–86.
[15] Schilling KG, Daducci A, Maier-Hein K, Poupon C, Houde JC, Nath V, et al. Challenges in diffusion MRI tractography–lessons learned from international benchmark competitions. Magn Reson Imaging 2019;57:194–209.
[16] Maier-Hein KH, Neher PF, Houde JC, Côté MA, Garyfallidis E, Zhong J, et al. The challenge of mapping the human connectome based on diffusion tractography. Nat Commun 2017;8(1):1349.
[17] Côté MA, Girard G, Boré A, Garyfallidis E, Houde JC, Descoteaux M. Tractometer: towards validation of tractography pipelines. Med Image Anal 2013;17(7):844–57.
[18] Girard G, Whittingstall K, Deriche R, Descoteaux M. Towards quantitative connectivity analysis: reducing tractography biases. Neuroimage 2014;98:266–78.
[19] Neher PF, Côté MA, Houde JC, Descoteaux M, Maier-Hein KH. Fiber tractography using machine learning. NeuroImage 2017 sep;158:417–29.
[20] Duru DG, Ozkan M. SOM based diffusion tensor MR analysis. Image and signal processing and analysis, 2007. ISPA 2007. 5th international symposium on IEEE. 2007. p. 403–6.
[21] Duru DG, Ozkan M. Self-organizing maps for brain tractography in MRI. Neural engineering (NER), 2013 6th international IEEE/EMBS conference on IEEE. 2013. p. 1509–12.
[22] Neher PF, Götz M, Norajitra T, Weber C, Maier-Hein KH. A machine learning based approach to Fiber Tractography using classifier voting. Cham: Springer; 2015. p. 45–52.
[23] Jörgens D, Smedby Ö, Moreno R. Learning a single step of streamline Tractography based on neural networks. Computational diffusion MRI. Springer; 2018. p. 103–16.
[24] Wegmayr V, Giuliari G, Holdener S, Buhmann J. Data-driven fiber tractography with neural networks. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) IEEE. 2018. p. 1030–3.
[25] Poulin P, Côté MA, Houde JC, Petit L, Neher PF, Maier-Hein KH, et al. Learn to track: Deep learning for Tractography. Cham: Springer; 2017. p. 540–7.
[26] Benou I, Riklin-Raviv T. DeepTract: A probabilistic deep learning framework for White matter Fiber Tractography arXiv preprint arXiv:181205129 2018.
[27] Poulin P, Rheault F, St-Onge E, Jodoin PM, Descoteaux M. Bundle-wise deep tracker: Learning to track bundle-specific streamline paths. Proceedings of the International Society for Magnetic Resonance in medicine ISMRM-ESMRMB. 2018.
[28] Wasserthal J, Neher PF, Maier-Hein KH. Tract orientation mapping for bundle-specific tractography. International conference on medical image computing and computer-assisted intervention. Springer; 2018. p. 36–44.
[29] OASd Lucena. Deep learning for brain analysis in MR imaging. SÃ£o Paulo, Brazil: [sn]. http://repositorio.unicamp.br/jspui/handle/REPOSIP/332646; 2018.
[30] Reisert M, Coenen VA, Kaller C, Egger K, Skibbe H. HAMLET: Hierarchical

harmonic filters for learning tracts from diffusion MRI arXiv preprint arXiv:180701068 2018.
[31] Wasserthal J, Neher P, Maier-Hein KH. TractSeg-fast and accurate white matter tract segmentation. NeuroImage 2018;183:239–53.
[32] Koppers S, Merhof D. Direct estimation of Fiber orientations using deep learning in diffusion imaging. Cham: Springer; 2016. p. 53–60.
[33] Ngattai Lam PD, Belhomme G, Ferrall J, Patterson B, Styner M, Prieto JC. TRAFIC: Fiber tract classification using deep learning. Proc SPIE Int Soc Opt Eng 2018 feb;10574.
[34] Patil SM, Nigam A, Bhavsar A, Chattopadhyay C. Siamese LSTM based Fiber structural similarity network (FS2Net) for rotation invariant brain Tractography segmentation undefined 2017.
[35] Gupta V, Thomopoulos SI, Rashid FM, Thompson PM. FiberNET: An ensemble deep learning framework for clustering White matter fibers. Cham: Springer; 2017. p. 548–55.
[36] Gupta V, Thomopoulos SI, Corbin CK, Rashid F, Thompson PM. FIBERNET 2.0: An automatic neural network based tool for clustering white matter fibers in the brain. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) IEEE. 2018. p. 708–11.
[37] Thomas C, Frank QY, Irfanoglu MO, Modi P, Saleem KS, Leopold DA, et al. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. Proc Natl Acad Sci 2014;111(46):16574–9.
[38] Pujol S, Wells W, Pierpaoli C, Brun C, Gee J, Cheng G, et al. The DTI challenge: toward standardized evaluation of diffusion tensor imaging tractography for neurosurgery. J Neuroimaging 2015;25(6):875–82.
[39] Yeh FC, Panesar S, Fernandes D, Meola A, Yoshino M, Fernandez-Miranda JC, et al. Population-averaged atlas of the macroscale human structural connectome and its network topology. NeuroImage 2018;178:57–68.
[40] Zhang F, Wu Y, Norton I, Rigolo L, Rathi Y, Makris N, et al. An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan. NeuroImage 2018;179:429–47.
[41] Thon A, Teichgräber U, Tennstedt-Schenk C, Hadjidemetriou S, Winzler S, Malich A, et al. Computer aided detection in prostate cancer diagnostics: a promising alternative to biopsy? A retrospective study from 104 lesions with histological ground truth. PLOS ONE 2017;12(10):1–21. 10.
[42] Clinic C. Alzheimer's disease: overview of diagnostic tests [Online; accessed 3-January-2019]. https://my.clevelandclinic.org/health/diagnostics/9176-alzheimers-disease-overview-of-diagnostic-tests/; 2014.
[43] Bernard O, Bosch JG, Heyde B, Alessandrini M, Barbosa D, Camarasu-Pop S, et al. Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. IEEE Trans Med Imaging 2016;35(4):967–77.
[44] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993.
[45] Fillard P, Descoteaux M, Goh A, Gouttard S, Jeurissen B, Malcolm J, et al. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. Neuroimage 2011;56(1):220–34.
[46] Wilkins B, Lee N, Singh M. Development and evaluation of a simulated FiberCup phantom. International symposium on magnetic resonance in medicine (ISMRMâ €™12). 2012. p. 1938.
[47] Yendiki A, Panneck P, Srinivasan P, Stevens A, Zöllei L, Augustinack J, et al. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. Front Neuroinform 2011;5:23.
[48] Daducci A, Canales-Rodríguez EJ, Descoteaux M, Garyfallidis E, Gur Y, Lin YC, et al. Quantitative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI. IEEE Trans Med Imaging 2014;33:384–99. EPFL-ARTICLE-183667.
[49] Daducci A, Caruyer E, Descoteaux M, Houde J, Thiran J. HARDI reconstruction challenge 2013. Proceedings of the IEEE international symposium on biomedical imaging (ISBI), San Francisco, CA. 2013.
[50] Chenot Q, Tzourio-Mazoyer N, Rheault F, Descoteaux M, Crivello F, Zago L, et al. A population-based atlas of the human pyramidal tract in 410 healthy participants. Brain Struct Funct 2018:1–14.
[51] Rheault F, St-Onge E, Sidhu J, Maier-Hein K, Tzourio-Mazoyer N, Petit L, et al. Bundle-specific tractography with incorporated anatomical and orientational priors. NeuroImage 2019;186:382–98.
[52] Poupon C, Laribiere L, Tournier G, Bernard J, Fournier D, Fillard P, et al. A diffusion hardware phantom looking like a coronal brain slice. Proceedings of the International Society for Magnetic Resonance in medicine. vol. 18. 2010. p. 581.
[53] Neher PF, Laun FB, Stieltjes B, Maier-Hein KH. Fiberfox: facilitating the creation of realistic white matter software phantoms. Magn Reson Med 2014;72(5):1460–70.
[54] White T, Magnotta VA, Bockholt HJ, Williams S, Wallace S, Ehrlich S, et al. Global white matter abnormalities in schizophrenia: a multisite diffusion tensor imaging study. Schizophr Bull 2009;37(1):222–32.
[55] Garyfallidis E, Côté MA, Rheault F, Sidhu J, Hau J, Petit L, et al. Recognition of white matter bundles using local and global streamline-based registration and clustering. NeuroImage 2018;170:283–95.
[56] Mazoyer B, Mellet E, Perchey G, Zago L, Crivello F, Jobard G, et al. BIL&GIN: a neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. Neuroimage 2016;124:1225–31.
[57] Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, et al. The WU-Minn human connectome project: an overview. Neuroimage 2013;80:62–79.
[58] Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, et al. The minimal preprocessing pipelines for the human connectome project. Neuroimage 2013;80:105–24.
[59] Wassermann D, Makris N, Rathi Y, Shenton M, Kikinis R, Kubicki M, et al. The white matter query language: a novel approach for describing human white matter anatomy. Brain Struct Funct 2016;221(9):4705–21.
[60] Stieltjes B, Brunner RM, Fritzsche K, Laun F. Diffusion tensor imaging: Introduction and atlas. Springer Science & Business Media; 2013.
[61] Garyfallidis E, Brett M, Correia MM, Williams GB, Nimmo-Smith I. Quickbundles, a method for tractography simplification. Front Neurosci 2012;6:175.
[62] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15(1):29.
[63] Reddy CP, Rathi Y. Joint multi-fiber NODDI parameter estimation and tractography using the unscented information filter. Front Neurosci 2016;10:166.
[64] MITK, MITK diffusion imaging; 2018. [Online; accessed 3-January-2019]. http://www.mitk.org/wiki/DiffusionImaging.
[65] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention (MICCAI), vol. 9351 of LNCS. 2015. p. 234–41.
[66] Manjón JV, Coupé P, Concha L, Buades A, Collins DL, Robles M. Diffusion weighted image denoising using overcomplete local PCA. PLoS One 2013;8(9):e73021.
[67] Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012;6(4):15.
[68] Aganj I, Lenglet C. Sapiro G. ODF reconstruction in q-ball imaging with solid angle consideration. Biomedical imaging: From Nano to macro, 2009. ISBI'09. IEEE international symposium on IEEE. 2009. p. 1398–401.
[69] Raudys SJ, Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans Pattern Anal Mach Intell 1991(3):252–64.
[70] Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.
[71] Cárdenas-Blanco A, Tejos C, Irarrazaval P, Cameron I. Noise in magnitude magnetic resonance images. Concepts Magn Reson Part A: An Educ J 2008;32(6):409–16.
[72] Varadarajan D, Haldar JP. A majorize-minimize framework for Rician and non-central chi MR images. IEEE Trans Med Imaging 2015;34(10):2191–202.
[73] Basu S, Fletcher T, Whitaker R. Rician noise removal in diffusion tensor MRI. International conference on medical image computing and computer-assisted intervention. Springer; 2006. p. 117–25.
[74] Kleesiek J, Petersen J, Döring M, Maier-Hein K, Köthe U, Wick W, et al. Virtual raters for reproducible and objective assessments in radiology. Sci Rep 2016;6:25007.
[75] Entis JJ, Doerga P, Barrett LF, Dickerson BC. A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution MRI. Neuroimage 2012;60(2):1226–35.
[76] Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, et al. Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance. Alzheimers Dement 2015;11(2):126–38.
[77] Piccinini F, Tesei A, Paganelli G, Zoli W, Bevilacqua A. Improving reliability of live/dead cell counting through automated image mosaicing. Comput Methods Programs Biomed 2014;117(3):448–63.
[78] Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta 2013;760:25–33.