# Supervised White Matter Bundle Segmentation in Glioma Patients with Transfer Learning *

Chiara Riccardi[1,2], Sofia Ghezzi[2], Gabriele Amorosino[1,2], Luca Zigiotto[3], Silvio Sarubbo[3], Jorge Jovicich[2], and Paolo Avesani[1,2]

[1] Neuroinformatics Laboratory, Fondazione Bruno Kessler, Italy
[2] Center for Mind/Brain Sciences, University of Trento, Italy
[3] Neurosurgery Department, Azienda Provinciale Servizi Sanitari, Italy
criccardi@fbk.eu avesani@fbk.eu

**Abstract.** In clinical neuroscience, the segmentation of the main white matter bundles is a mandatory premise for pre-operative neurosurgical planning, neuro-rehabilitation prognosis and monitoring of neuro-related diseases. The automation of the task of bundle segmentation is achieving a good accuracy on healthy individuals taking advantage of data driven approaches and deep learning models. The lack of large clinical datasets is preventing the translation of these results to patients. Inference on patients data with model trained on healthy population is not effective because of domain shift. This study aims to carry out an empirical analysis to investigate how techniques of transfer learning might be beneficial to overcome these limitations. For our analysis, we consider an open publicly available dataset with hundreds of individuals and a clinical dataset with several tens of glioma patients. We focus our preliminary investigation on corticospinal tract. The results show that transfer learning might be effective to overcome the component of domain shift related to systemic bias. Nevertheless, the alterations induced by the tumors are not successfully managed by transfer learning and are still preventing the successful translation of model trained on healthy population.

**Keywords:** White Matter Bundle Segmentation · Glioma · Deep learning · Transfer learning.

## 1 Introduction

This work aims to carry out an empirical study to investigate how effective might be the translation of a supervised learning model for bundle segmentation from a healthy population to patients with glioma.

In clinical neuroscience the identification of the main white matter bundles is a mandatory premise for pre-operative neurosurgical planning, the prognosis of neuro-rehabilitation treatment and more in general to monitor the progression

* This work was partially supported by the grant PAT Reg. n. 764/2021 NeuSurPlan. We also acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU

of neuro-related diseases [5]. In particular, for patients with glioma, the segmentation of bundles provides to the neurosurgeons a reference for the eloquent regions to be preserved in order to reduce subsequent cognitive impairment [18].

Manual bundle segmentation is time-consuming. For this reason, the research community spent a meaningful effort to develop automated methods to carry out this task [5]. Data-driven strategy and deep learning models achieved a meaningful performance on volumetric representation of the white matter structure [16, 4, 11, 9, 19]. A key factor for this result is the availability of large open datasets with the annotated bundles [16, 13].

The lack of analogous large open clinical datasets is preventing the training of the supervised learning model tailored for patients, where the white matter structures deviate from the normative model. Despite the occasional use of the models trained on healthy individuals for some small collection of patients [14, 17, 10, 2, 12], the translation of these models to clinical data remains an open challenge.

The source of complexity is represented by the deviation of patient data from the distribution of healthy population, and consequently the mismatch between training and inference data. Such a kind of mismatch is known in literature as domain shift. The factors that contribute to domain shift might be manifolds: the presence of a tumor, the bias of the operator, the methods of DWI processing, the quality of data. The usual techniques to deal with domain shift are commonly referred to as transfer learning [6].

Our empirical study is addressing some open research questions such as: what might be the drop in performance carrying out inference on patients using the model trained on healthy individuals? Might a small set of clinical data be enough to successfully retrain the model? Does a technique of transfer learning provide an additional improvement in the accuracy of bundle segmentation for patients?

We focus our preliminary analysis on the cortical spinal tract (CST) since the anatomical definition is less controversial among neuroanatomists [3]. The learning of the normative model of CST is carried out by referring to an open publicly available dataset of hundreds of healthy individuals with manual annotation of the bundles. The testing, the retraining, and the transfer learning are performed on a collection of 21 patients with glioma.

The results show that the drop in translating a trained model from healthy individuals to patients is almost of 20%. Transfer learning is effective in partially recovering such a gap. Nevertheless, only the portion of domain shift related to systemic bias is corrected while the portion of errors due to the alterations of the tumors remains an open challenge.

## 2 Materials

The data-driven investigation is designed upon two datasets: an open publicly available dataset of healthy individuals and a clinical dataset of patients with

gliomas. For each individual in both datasets, we refer to the T1w brain images and the annotation of left CST volumetric masks.

Data of healthy individuals are drawn from TractoInferno [13], a multisite collection of hundreds of MR recordings, both T1w and DWI, and their derivatives such as the constrained spherical deconvolution (CSD) diffusivity model, the fiber orientation distribution function (fODF), and several manual annotations of main white matter bundles, extracted from ensemble tractography. We consider only the 242 individuals where the experts find a common agreement of left CST identification.

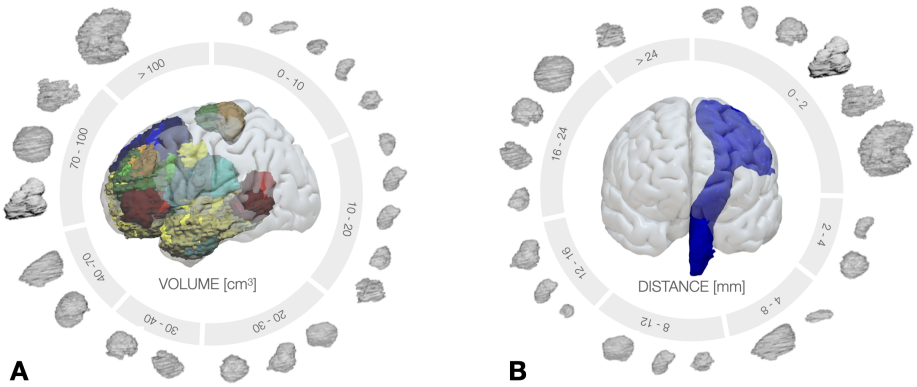


**Fig. 1.** Clinical dataset: (A) spatial distribution of tumors in the MNI space and their volumes reported in $cm^3$, (B) distance distribution of the tumors from the left CST of each patient in $mm$.

Data of patients with glioma are acquired by the Neurosurgery Department at the S. Chiara Hospital of Trento. Details on the acquisition and processing of data are described in [20]. We selected 21 patients with tumors located in the left hemisphere, as illustrated in Figure 1. For each patient we consider the following processing: (i) the manual segmentation of the region of tumor reporting the volumetric mask onto the T1w image, (ii) the CSD diffusivity model reconstruction, (iii) the probabilistic tractography, (iv) the manual segmentation of CST and the related volumetric masks. All data are registered to MNI space with a linear transformation.

The two datasets are encoding a twofold domain shift. The former is represented by the alteration of the left CST due to the presence of tumors and the consequent deviation from the normative model. The latter refers to a systemic bias composed of many factors such as the acquisition parameters, the tractography computation, and the manual annotation of bundles.

## 3 Methods

### 3.1 The supervised learning models

Although there are many different approaches to automate models for white matter segmentation in literature, we focus our analysis on methods based on volumetric representation. This choice aims to reduce the additional variance of bundle representation derived from tractography. The reference method for volumetric representation of bundle segmentation is a 3D U-Net deep learning model [16]. The choice of the input to the model can span from the scalar intensity of T1w image, to different DWI data derivatives such as fractional anisotropy (FA) or the main peak of fiber orientation distribution function (fODF). A recent work [19] shows that T1w can be quite effective in capturing the spatial distribution of bundles and a viable option when DWI data might be not available such as in the clinical practice. For these reasons, we design our analysis considering T1w images as reference data for bundle segmentation.

For our investigation we consider the 3D U-Net as reference learning model but with three different training strategies. The first strategy is designed as the default training process where the weights of the model are randomly initialized. The second strategy resumes the previous training process with a new set of data, and in the following it is referred to as *r*etrain. Network weights are mutuated by the last epoch while model hyperparameters are optimized with respect to the new training data. The third strategy is transfer learning [15]. It is analogous to retrain strategy but with a different schema of weights update. Learning is restricted to only a portion of the encoder-decoder network, the second half in charge of reconstructing the original space from latent space [15].

### 3.2 Performance evaluation

To quantify the performance of the learning models after the different training strategies we use two measures. The first is the Dice Score Coefficient (DSC) that evaluates voxel-wise the global similarity between the predicted bundle mask and the related ground truth defined by manual annotation. The second metric aims to support a pairwise comparison of two training strategies at the voxel level when we operate the inference on the same collection of data. The goal is to evaluate the spatial distribution of the regions where one training strategy outperforms the other. In the following, we refer to this second measure as *Improvement Map*.

The computation of the Improvement Map takes in input a set of individuals and for each of them, two bundle masks generated by two distinct models as prediction of true bundle segmentation. Each of these bundles mask is converted into a binary true prediction map. This binary map has values 1 in voxels where either negative or positive true predictions occur, 0 otherwise. An individual Improvement Map is obtained by computing the difference between the true prediction maps derived by the two models. When the behavior of the two models is the same, the voxels will have value 0, 1 or $-1$ otherwise. We may

obtain a global Improvement Map by averaging the individual maps, across the population. Given two competing models, positive values of the Improvement Map account for better performance of the former model, negative values for the latter model. Values close to 0 denote similar behavior.

### 3.3 Domain shift assessment

The characterization of the domain shift is carried out following two perspectives: one oriented to the volumetric representation and another to the tractography representation. The first is considering the spatial distribution of the bundle masks. The results are two probabilistic maps, one for healthy individuals, the other for glioma patients. The second measure refers to the main pathway of a bundle, that we represent with the most representative fiber, sometimes referred to as the *skeleton* [8] of the bundle. The intuitive idea is to measure the deformation of the skeleton with respect to the normative model. For this purpose, we introduce the notion of *Warp V-Norm* and *Warp V-Norm profile*.

The Warp V-Norm estimates the local shape differences between two bundles, encoded as volumetric binary masks after a spatial normalization with an affine registration in a common reference space. The first step is the computation of the warp by registering the two bundles, combining a rigid and a diffeomorphic transformation. The Warp V-Norm is defined as the 2-norm of the vector encoding the whole transformation in each voxel of the warped image. High values denote a major displacement or alteration of a given bundle with respect to the normative model.

The Warp V-Norm profile aims to summarize the volumetric information of deformation with respect to the skeleton of the normative model. The reference skeleton is derived from the fiber representation of all the bundles of healthy population, following the process described in [1]. The skeleton profile is a vector where each point of the fiber is associated to the Warp V-Norm of the correspondent voxel.

## 4 Experiments and Results

The experiments are designed to assess the performance of the supervised learning models trained according to the three reference strategies: no transfer learning, model retraining, transfer learning. The additional goal is to characterize the behavior of the different learning strategies with respect to the different factors of domain shift between healthy and patients.

We first focus on how to measure the domain shift between the two datasets considering the volumetric representation of the bundles as voxel masks. Tractography representation of bundles is first mapped to voxels, then smoothed and registered with an affine transformation to the MNI common reference space. Separately for the two datasets, the spatial distribution of the bundles is computed as the relative frequency of each voxel to be part of the CST. To highlight

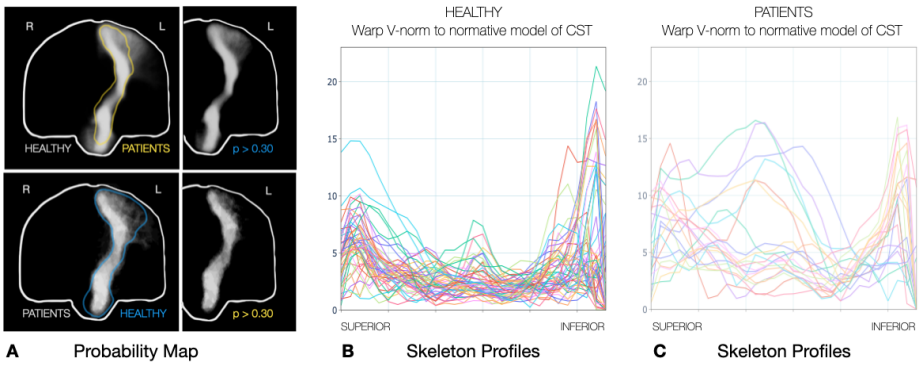


**Fig. 2.** CST normative model and clinical deviation: (A) the probability map of the volumetric masks for healthy and patients, respectively top and bottom, and their envelope for $p > 0.30$; (B) the Warp V-Norm profile of healthy test set with respect to normative model along the skeleton pathway; (C) similar to the previous panel but for patients.

the difference between the two maps we provide a thresholded version ($p > 0.30$) in Figure 2.A.

For characterizing the domain shift with respect to the tractography representation, we compute the Warp V-Norm profile for 48 individuals from the healthy population (those selected as healthy test set) and for the 21 glioma patients. The template of bundle mask for the normative model is derived from the probability map of healthy individuals, thresholded so that its volume matches the mean volume of the whole population. Rigid and diffeomorphic registrations of each individual's bundle with respect to the template are computed using ANTs. The displacement along the skeleton for each individual with respect to the normative model is separately reported in Figure 2.B and 2.C, healthy and patients respectively.

For reducing the computational overload and to balance the ratio between the classes of the binary classifiers we apply a heuristic to resize the bounding box of T1w images, by excluding the controlateral hemisphere, the occipital, and frontal lobes. We design 3 training sessions and 6 inference sessions. In the first session a 3D U-Net model is trained with healthy data (M-train), in the second the M-train model is retrained with patients data only (M-retrain), in the third the M-train model is trained with a transfer learning method with patients data only (M-transfer). Each of these three models is used for inference both on healthy and patient test sets. Performance is measured as average DSC between predicted and true bundle masks. Results are reported in Table 1.

Train and test split is designed differently for healthy and patients data. The former is divided with a ratio of 80/20, with 193 individuals for training and 48 for test. Hyperparameters' learning is carried out with a three fold cross-validation schema and 800 epochs each. Differently for patients, where the sample size is small, we follow a stratified cross-validation schema. Data are partitioned

according to a 5-fold cross validation, and a nested 3-fold partition for hyperparameters' estimation is performed with 200 epochs each. Variance of segmented bundles is much larger in patients, for this reason, the design of partitioning is revised to balance the distribution of bundles with major alterations. As measure of alteration, we refer to the Warp V-Norm profile computed as reported above.

As reference implementation of 3D U-Net, we refer to nnU-Net [7], a state-of-the art deep-learning framework for biomedical image segmentation.

**Table 1.** Learning assessment. Mean DSC and standard deviations of the inference on healthy and patient datasets when the 3D U-Net model is trained with the three different training strategies.

| | M-train (Healthy) | M-retrain (Patient) | M-transfer (Patient) |
|---|---|---|---|
| Healthy dataset | $0.83 \pm 0.025$ | $0.67 \pm 0.049$ | $0.73 \pm 0.040$ |
| Patients dataset | $0.68 \pm 0.060$ | $0.75 \pm 0.101$ | $0.74 \pm 0.097$ |

Pairwise comparison of models' performance is carried out at the voxel level, by computing the Improvement Maps, looking at the True Positive and True Negative errors, as illustrated above. We focus our analysis on the contrast between M-transfer and M-train models when inference is operated on patients. The improvement map of True Positive and True Negative reports the spatial distribution where M-transfer outperforms M-train. Figure 3.A shows the portion of domain shift successfully managed by a strategy of transfer learning. Similarly, we analyze the contrast between M-transfer and M-retrain models when inference is operated on the healthy test set, see Figure 3.B.

## 5 Discussion and Conclusions

Results reported in Table 1 provide the empirical evidence that there is a meaningful drop in performance when the model trained on TractoInferno is not applied to healthy population, but to patients' data, 0.83 versus 0.68.

The reasons for this loss in accuracy are depicted in Figure 2.A. There is a clear mismatch between the probability maps of CST voxel masks of the two populations. The differences are located mainly in the inferior and the superior portion of the bundles. This component of domain shift seems to be related to a systemic bias in the patients' data, partially concerned with tractography reconstruction and partially due to the manual annotation of different experts. The skeleton Warp V-Norm profiles for the two populations, as shown in Figures 2.B and 2.C, provide a visual representation of the domain shift due to the alterations induced by the tumors. The main displacement of CST pathways is in the region of internal capsule probably related to the tumors located in the temporal lobe.

Both retraining and transfer learning look to be effective in partially recovering the domain shift, 0.68 versus 0.74. The models capture the different

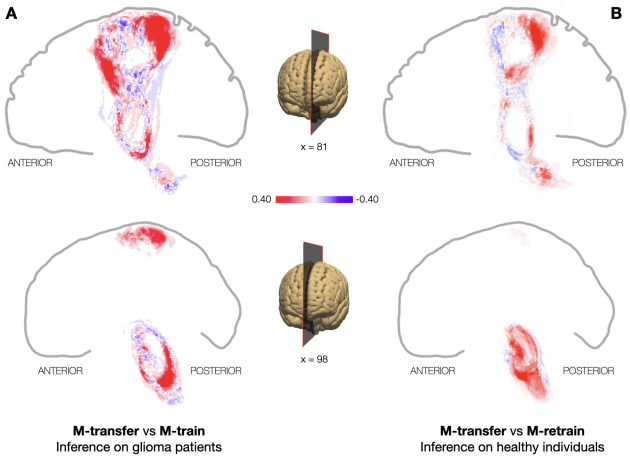


**Fig. 3.** Transfer learning Improvement Maps. (A) A voxelwise analysis to contrast the behavior of M-train model with respect to M-transfer when used for inference on patients data. Red color highlights the region where M-transfer outperforms M-train, blue color otherwise. (B) Contrast analysis between M-transfer and M-retrain when used for inference on healthy population. In both panels a snapshot of Improvement Maps for two sagittal planes, $x = 81$ and $x = 98$ of the MNI template.

distribution of patients' data. It is worthwhile to remark that we achieve these results with only a 10% portion of data with respect to those used for training the original normative model.

While inference on patients' data are substantially equivalent for both retraining and transfer learning models, there is a meaningful difference if we focus on the inference of healthy population, 0.67 versus 0.73. As expected, transfer learning better preserves the balance between the two populations.

The empirical results show only a partial recovery of the domain shift. There is still a large margin of improvement to achieve a performance in patients similar to healthy. The analysis of recovery reported in Figure 3 shows the regions of CST masks where the accuracy increases after the transfer learning. These regions are located in the inferior area of the stem and in the posterior part of the parietal terminations of CST. These differences seem related to systemic bias of patients' data: (i) lack of inferior slices in MR acquisition for patients, and (ii) tractography reconstruction in the superior part of CST. Apparently, there is no improvement in the middle part where, as shown in Figure 2.C, we have the major deviation from normative model due to tumors.

Conclusions of this study suggest that transfer learning might be an effective method to manage the systemic bias of domain shift and at the same time an efficient way for exploiting the small sample size of clinical data. Nevertheless, the segmentation of bundles with meaningful deviations from normative model requires a different strategy of learning.

## References


1. Amorosino, G., Olivetti, E., Jovicich, J., Avesani, P.: How Does White Matter Registration Affect Tractography Alignment? preprint, NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation, Trento, Italy; Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy (Oct 2022). https://doi.org/10.1101/2022.10.27.513915, http://biorxiv.org/lookup/doi/10.1101/2022.10.27.513915
2. Bertò, G., Bullock, D., Astolfi, P., Hayashi, S., Zigiotto, L., Annicchiarico, L., Corsini, F., De Benedictis, A., Sarubbo, S., Pestilli, F., Avesani, P., Olivetti, E.: Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. NeuroImage **224**, 117402 (Jan 2021). https://doi.org/10.1016/j.neuroimage.2020.117402, https://www.sciencedirect.com/science/article/pii/S1053811920308879
3. Chenot, Q., Tzourio-Mazoyer, N., Rheault, F., Descoteaux, M., Crivello, F., Zago, L., Mellet, E., Jobard, G., Joliot, M., Mazoyer, B., Petit, L.: A population-based atlas of the human pyramidal tract in 410 healthy participants. Brain Struct Funct **224**(2), 599–612 (Mar 2019). https://doi.org/10.1007/s00429-018-1798-7, https://doi.org/10.1007/s00429-018-1798-7
4. Dong, X., Yang, Z., Peng, J., Wu, X.: Multimodality white matter tract segmentation using CNN. In: Proceedings of the ACM Turing Celebration Conference - China. pp. 1–8. ACM TURC '19, Association for Computing Machinery, New York, NY, USA (May 2019). https://doi.org/10.1145/3321408.3326673, https://dl.acm.org/doi/10.1145/3321408.3326673
5. Ghazi, N., Aarabi, M.H., Soltanian-Zadeh, H.: Deep Learning Methods for Identification of White Matter Fiber Tracts: Review of State-of-the-Art and Future Prospective. Neuroinform (Jun 2023). https://doi.org/10.1007/s12021-023-09636-4, https://doi.org/10.1007/s12021-023-09636-4
6. Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., Azim, M.A.: Transfer learning: a friendly introduction. Journal of Big Data **9**(1), 102 (Oct 2022). https://doi.org/10.1186/s40537-022-00652-w, https://doi.org/10.1186/s40537-022-00652-w
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, https://www.nature.com/articles/s41592-020-01008-z, number: 2 Publisher: Nature Publishing Group
8. Li, W., Hu, X.: Robust Tract Skeleton Extraction of Cingulum Based on Active Contour Model from Diffusion Tensor MR Imaging. PLoS ONE **8**(2), e56113 (Feb 2013). https://doi.org/10.1371/journal.pone.0056113, https://dx.plos.org/10.1371/journal.pone.0056113
9. Lu, Q., Li, Y., Ye, C.: Volumetric white matter tract segmentation with nested self-supervised learning using sequential pretext tasks. Medical Image Analysis **72**, 102094 (Aug 2021). https://doi.org/10.1016/j.media.2021.102094, https://www.sciencedirect.com/science/article/pii/S1361841521001407
10. Moshe, Y.H., Ben Bashat, D., Hananis, Z., Teicher, M., Artzi, M.: Utilizing the TractSeg Tool for Automatic Corticospinal Tract Segmentation in Patients With Brain Pathology. Technol Cancer Res Treat **21**, 15330338221131387 (Jan 2022). https://doi.org/10.1177/15330338221131387, https://doi.org/10.1177/15330338221131387, publisher: SAGE Publications Inc

11. Nelkenbaum, I., Tsarfaty, G., Kiryati, N., Konen, E., Mayer, A.: Automatic Segmentation of White Matter Tracts Using Multiple Brain MRI Sequences. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 368–371 (Apr 2020). https://doi.org/10.1109/ISBI45749.2020.9098454, iSSN: 1945-8452
12. O'Donnell, L.J., Suter, Y., Rigolo, L., Kahali, P., Zhang, F., Norton, I., Albi, A., Olubiyi, O., Meola, A., Essayed, W.I., Unadkat, P., Ciris, P.A., Wells, W.M., Rathi, Y., Westin, C.F., Golby, A.J.: Automated white matter fiber tract identification in patients with brain tumors. NeuroImage: Clinical **13**, 138–153 (Jan 2017). https://doi.org/10.1016/j.nicl.2016.11.023, https://www.sciencedirect.com/science/article/pii/S2213158216302315
13. Poulin, P., Theaud, G., Rheault, F., St-Onge, E., Bore, A., Renauld, E., de Beaumont, L., Guay, S., Jodoin, P.M., Descoteaux, M.: TractoInferno - A large-scale, open-source, multi-site database for machine learning dMRI tractography. Sci Data **9**(1), 725 (Nov 2022). https://doi.org/10.1038/s41597-022-01833-1, https://www.nature.com/articles/s41597-022-01833-1, number: 1 Publisher: Nature Publishing Group
14. Richards, T.J., Anderson, K.L., Anderson, J.S.: "Fully automated segmentation of the corticospinal tract using the TractSeg algorithm in patients with brain tumors". Clinical Neurology and Neurosurgery **210**, 107001 (Nov 2021). https://doi.org/10.1016/j.clineuro.2021.107001, https://www.sciencedirect.com/science/article/pii/S0303846721005308
15. Sundaresan, V., Zamboni, G., Dinsdale, N.K., Rothwell, P.M., Griffanti, L., Jenkinson, M.: Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain MR images. Medical Image Analysis **74**, 102215 (Dec 2021). https://doi.org/10.1016/j.media.2021.102215, https://www.sciencedirect.com/science/article/pii/S1361841521002607
16. Wasserthal, J., Neher, P., Maier-Hein, K.H.: TractSeg - Fast and accurate white matter tract segmentation. NeuroImage **183**, 239–253 (Dec 2018). https://doi.org/10.1016/j.neuroimage.2018.07.070, https://www.sciencedirect.com/science/article/pii/S1053811918306864
17. Xue, T., Zhang, F., Zhang, C., Chen, Y., Song, Y., Golby, A.J., Makris, N., Rathi, Y., Cai, W., O'Donnell, L.J.: Superficial white matter analysis: An efficient point-cloud-based deep learning framework with supervised contrastive learning for consistent tractography parcellation across populations and dMRI acquisitions. Medical Image Analysis **85**, 102759 (Apr 2023). https://doi.org/10.1016/j.media.2023.102759, https://www.sciencedirect.com/science/article/pii/S1361841523000208
18. Yang, J.Y.M., Yeh, C.H., Poupon, C., Calamante, F.: Diffusion MRI tractography for neurosurgery: the basics, current state, technical reliability and challenges. Phys. Med. Biol. **66**(15), 15TR01 (Jul 2021). https://doi.org/10.1088/1361-6560/ac0d90, https://dx.doi.org/10.1088/1361-6560/ac0d90, publisher: IOP Publishing
19. Yang, Q., Hansen, C.B., Cai, L.Y., Rheault, F., Lee, H.H., Bao, S., Chandio, B.Q., Williams, O., Resnick, S.M., Garyfallidis, E., Anderson, A.W., Descoteaux, M., Schilling, K.G., Landman, B.A.: Learning white matter subject-specific segmentation from structural MRI. Med Phys **49**(4), 2502–2513 (Apr 2022). https://doi.org/10.1002/mp.15495
20. Zigiotto, L., Vavassori, L., Annicchiarico, L., Corsini, F., Avesani, P., Rozzanigo, U., Sarubbo, S., Papagno, C.: Segregated circuits for phonemic and semantic fluency: A novel patient-tailored disconnection study. NeuroIm-

age: Clinical **36**, 103149 (Jan 2022). https://doi.org/10.1016/j.nicl.2022.103149, https://www.sciencedirect.com/science/article/pii/S2213158222002145