PRELIMINARY RESULTS ON THE USE OF STAPLE FOR EVALUATING DT-MRI TRACTOGRAPHY IN THE ABSENCE OF GROUND TRUTH

S. Pujol[1], C-F. Westin[2], R. Whitaker[3], G. Gerig[3], T. Fletcher[3], V. Magnotta[4], S. Bouix[5], R. Kikinis[1], W. M. Wells III[1], and R. Gollub[6]

[1]Surgical Planning Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [2]Laboratory of Mathematics in Imaging, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, [3]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah, [4]University of Iowa, Iowa City, IA, [5]Psychiatry Neuroimaging Laboratory, Boston, MA, [6]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA

Introduction DT-MRI Tractography provides a non-invasive approach for reconstructing the 3D trajectory of white matter fasciculi. The potential for inferring the connectivity of the brain *in-vivo* would bring valuable non-invasive resources for scientific study or therapeutic decisions. The lack of ground truth remains a persistent challenge for evaluating fiber tracking, and validation studies using synthetic phantoms and histological samples have been attempted. However, while synthetic data do not represent the physiological conditions of the brain, in-vitro sample preparation introduces geometric distortion that can alter tissue microstructure. Manual segmentation of tracts from scalar or DTI is difficult, and subject to inter-rater variation. Assessing the uncertainty in fiber orientation and data artifacts can provide valuable information for interpreting DT-MRI tractography reconstructions [1]. We propose to use STAPLE [2] (Simultaneous Truth and Performance Level Estimation), an Expectation-Maximization (EM) algorithm, to compute a probabilistic estimate of the hidden true white matter pathways, and a measure of the performance levels parameters of a series of tractography results.

Materials and Methods: Given as input a set of segmentations, STAPLE was designed to provide maximum likelihood (ML) estimates of the *True Positive Frequency* or *Sensitivity (p),* and *True Negative Frequency* or *Specificity (q),* as well as a probabilistic estimate of the unknown true segmentation.

Digital Phantom Experiments. To demonstrate the utility of STAPLE for characterizing the quality of tract reconstruction, and for estimating the true segmentation, we prepared mathematical phantom DWI data, for which the true noise-free segmentation is available for testing purposes. We simulated a synthetic helical DWI phantom with 12 diffusion gradient directions, 1 baseline non-diffusion-weighted signal and b= 800 s/mm², and we defined 2 regions of interest in the fractional anisotropy map of the phantom image. We generated Rician Noise by adding to the synthetic dataset a random sample from a Gaussian distribution with mean of 0, and standard deviations set to simulate 6 SNR levels: low SNR (10-14 dB), medium SNR (20-28 dB) and high SNR (40-56 dB) corresponding to six different acquisition times (T/8,T/4,T/2,T,2T,4T) [3]. A streamline tractography algorithm was used to reconstruct fibers at all noise levels, and the resulting tracts were converted into voxelwise binary label maps.

Brain Image Experiments. We ran 4 different tractography algorithms A1) Fiber Tracking [5], A2) 3DSlicer [6], A3) Gtract [7], A4) Volumetric Connectivity [8] on a subject dataset acquired on a 1.5 T Siemens using 8-Channel Coil with 60 gradient directions and 10 baselines. The acquisition parameters were as follows: b = 700 s/mm², TR = 8900 ms, TE = 80 ms, FOV = 256 mm, 2.0 mm voxel size, 2 mm slice thickness, 128 x 128 x 64 matrix size, and the DWI data were post processed for eddy current and EPI distortion correction. Each algorithm performed tractography between pre-defined anatomical regions of interest (ROIs) according to known anatomical landmarks. We reconstructed 4 tracts of interest: the Internal Capsule (IC), the Corpus Callosum Forceps Minor and Major (CCF Maj. and Min.) and the Fornix (FNX), and we voxelized the sets of reconstructed fibers. We ran the STAPLE algorithm on the portions of the voxelized tracts that were strictly located between the two ROIs, to compute the sensitivity and specificity of each method along with an estimate of the true tracts.

Results and Discussion The results on synthetic data show that STAPLE was able to well characterize the quality of the input segmentations, despite *not* having access to the true segmentation. In general, the estimated (p,q) are within 10% of the true values, and the estimated quality of the segmentations, in terms of their sensitivity, decreases in a sensible way as the level of noise in the DTI increases. Note that in all cases the specificity is near one, which is a consequence of the relatively small volume occupied by the voxelized tracts.

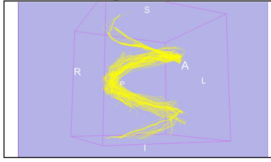
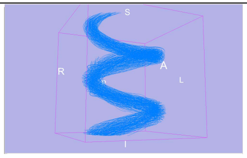
	Exact performance of tractography							
	SNR(dB)	10	14	20	28	40	56	
	(p,q)	(0.01, 0.99)	(0.08, 0.99)	(0.53, 0.99)	(0.84, 0.96)	(0.89, 0.99)	(0.98, 0.99)	
	STAPLE estimate performance of tractography							
	SNR(dB)	10	14	20	28	40	56	
	(p,q)	(0.01, 0.99)	(0.08, 0.99)	(0.57, 0.99)	(0.91, 0.96)	(0.97, 0.99)	(0.97, 0.99)	

Table 1 shows the sensitivity and specificity estimates with 0.01 precision that result from running STAPLE on the 6 segmentations corresponding to the synthetic noisy DTI. As the true (zero noise) segmentation is available for this mathematical phantom, we can calculate directly the "true" sensitivity and specificity of the six input segmentations. The figures show the helical tracts at SNR=20 dB (yellow) and SNR=56 dB (blue).

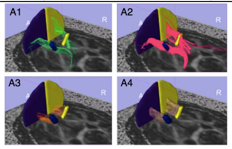
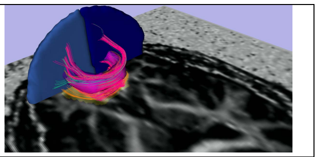
	Tract	**Left IC**	**Right IC**	**CC Maj.**	**CC Min.**	**Fornix**	
	A1 (p,q)	(0.52, 0.99)	(0.52, 0.99)	(0.18, 1.00)	(0.65, 0.99)	(0.15, 1.00)	
	A2 (p,q)	(0.80, 0.99)	(0.81, 0.99)	(0.55, 0.99)	(0.90, 0.99)	(0.62, 0.99)	
	A3 (p,q)	(0.42, 0.99)	(0.42, 0.99)	(0.37, 0.99)	(0.32, 0.99)	(0.51, 0.99)	
	A4 (p,q)	(0.74, 0.99)	(0.75, 0.99)	(0.91, 0.99)	(0.37, 0.99)	(0.52, 0.99)	

Table 2 shows the sensitivity and specificity values computed by STAPLE. The figures show the tractography results for the 4 algorithms A1 (green), A2 (pink), A3 (orange) and A4 (purple): in the IC (left figure), and in the CC. Min with the estimation of the true tract (yellow) overlaid on the Fractional Anisotropy image (right figure). The estimated performance level of the methods, in terms of sensitivity, varies across the four tracts – no single method dominates.

For the IC, A2 scores best, somewhat better than A4. For the CC examples, the scores are well separated, for CCF Maj., A4 is best and A2 next; for CCF Min. A2 is followed by A1. For the FNX, often considered a difficult structure, A2 is best, just somewhat better than A3 and A4. For most of the examples, A3 had the lowest scores, though it was competitive for FNX. A4 is based on volumetric connectivity and performs competitively with streamline algorithms. Overall, A2 scores most consistently well, and the brain data experiments results are consistent with visual inspection of the tracts. The relative lower performance of A3 may be due the relatively sparser collection of streamlines that it generated, which leads us to note the following: while the algorithms were run by their respective developers on data and ROIs that we supplied, it may be that the relative performance of a specific method could be improved by parameter tuning and other algorithmic adjustments. We plan to test the effect of using the truth estimate of the tracts to enable such optimization.

Conclusion This pilot study proposes a framework for validating DT-MRI tractography that has been designed for meeting the challenge that arises from the absence of ground truth. The synthetic data experiments demonstrated the precision of STAPLE in estimating the sensitivity and specificity of fiber tract reconstruction, and the results obtained on DT-MRI images of the brain provided preliminary assessment of the performances of four tractography algorithms, along with a statistical estimate of the ground truth estimation of four white matter tracts.

References [1] Jones et al. Magn Reson Med. 2005;53(6):1462-7 [2] Warfield SK et al. STAPLE. IEEE Trans Med Imaging. 23(7):903-21. [3] Chang et al. Magn Reson Med. 2007;57(1):141-13. [4] Wakana et al. Neuroimage 2007; 36(3):630-44 [5] Pierre Fillard, John Gilmore, Weili Lin, Joseph Piven, Guido Gerig, Quantitative Analysis of White Matter Fiber Properties along Geodesic Paths, Lecture Notes in Computer Science LNCS #2879, pp. 16-23, 2003 [6] www.slicer.org [7] Cheng P, Magnotta VA, Wu D, Nopoulos P, Moser DJ, Paulsen J, Jorge R, Andreasen NC. Evaluation of the GTRACT diffusion tensor tractography algorithm: a validation and reliability study. Neuroimage 2006. 31(3):1075-85.[8] PT Fletcher, R Tao, W-K Jeong, RT Whitaker. A volumetric approach to quantifying region-to-region white matter connectivity in diffusion tensor MRI, IPMI 2007, pp. 346-358.