Validation of automatic segmentation algorithms of DWI in acute stroke patients in independent data

Steven Mocking[1,2], Raquel Bezerra[1], Elissa McIntosh[1], Izzudin Diwan[1], Priya Garg[1], William Taylor Kimberly[3], Ethem Murat Arseva[1], Hakan Ay[1], Aneesh B Singhal[3], William A Copen[4], Pamela Schaefer[4], and Ona Wu[1]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Dept. of Radiology, MGH, Charlestown, MA, United States, [2]Image Sciences Institute, University Medical Center Utrecht, Utrecht, Netherlands, [3]Dept. of Neurology, MGH, Boston, MA, [4]Dept. of Radiology, MGH, Boston, MA

Background and purpose: Changes in diffusion-weighted MRI (DWI) are an early[1] and sensitive[2] marker in the diagnosis and prognosis of acute ischemic stroke. Accurate and automatic estimation of DWI lesions has application in clinical management as well as stroke research in that it ensures reproducibility of results. We evaluated five previously reported algorithms[3,4] for rapid automatic outlining of DWI lesions in an independent dataset for validating these methods. Automatically generated outlines were compared to manual delineations of acute imaging. Follow-up (FU) imaging was used to further investigate the outcome of tissue where disagreement between automatic and manual outline occurred for the acute time-point.

Methods: Cases of acute ischemic stroke patients who were imaged with DWI and perfusion-weighted imaging within 12h of last known well time, received neither intravenous nor intra-arterial recanalization therapy nor experimental therapy, had FU imaging ≥ 4 days and were not used in the training and testing of the original algorithms [3] were retrospectively analyzed (N=122). Trace diffusion-weighted MRI (DWI) was computed as the geometric mean of the high-b acquisitions [5]. The b_0 (b-value=0) image was used as the T2-weighted image (T2WI) and apparent diffusion coefficient (ADC) maps computed as the slope of the log of the DWI and the T2WI. Manual outlines of lesions on DWI mere made with knowledge of ADC.FU lesions, either FLAIR MRI or non-contrast CT were manually delineated (Display, MNI, McGill University) using visually determined thresholds. All outlines were made by a reader blinded to the output of the automatic algorithms. FU images were co-registered to the acute DWI (MNI Autoreg [6]). Acute DWI, ADC and T2WI were automatically segmented using five algorithms: ADC thresholding using ADC < $600 \cdot 10^{-6}$ s/mm^2[3,4,7], k-means[3,4], ISODATA[3,4,8], k-nearest neighbor[3,4,9] (k-NN) and Naive Bayes[4,10] (NB). NB computes the mean and variance of each feature for each class based on training data. It then labels sample voxels by assigning the class with the maximum likelihood. Likelihood is computed by estimating the probability of each feature having the observed value based on the trained mean and standard deviation of that feature; odds for all features are then multiplied together. The algorithms' performances were evaluated in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) with regard to the acute manual outline (AMO). Sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)) and Dice similarity coefficient (DSC, (2TP)/(2TP+FP+FN)) were calculated. To assess whether the misclassifications were due to subjective bias from manual outlines or poor algorithm performance, the FP voxels, where available, were also compared to the FU manual outline (FMO). The fraction of FP voxels outlined on FMO, the false FP (FFP), is computed. Similarly, false FM (FFN) is operationally defined as FN voxels not outlined on FMO. For comparing the classification methods, one-way ANOVA was performed with post-hoc two-sided Wilcoxon test. Holm-Bonferroni correction was used to adjust for multiple comparisons. Results are reported as mean±standard deviation or median [interquartile range].

Results: Examples of the occurrence of both FFP and FFN can be seen in Fig. 1. Sensitivity, specificity and DSC were significantly affected by choice of classifier ($p < 0.001$). Lesion volumes were 7.1 cm^3 [1.5-26.9] for AMO and 9.4 cm^3[1.9-44.5] for FMO . Results are provided in Table 1; performance of the algorithms on the training data is also provided for reference. NB performed significantly better than other approaches in terms of sensitivity ($p < 0.0001$) and DSC ($p < 0.05$), but lagged in specificity ($p < 0.01$). The fractions of FFP and FFN voxels can be seen in Table 2. There was no significant difference in %FFN between classifiers ($p = 0.38$), whereas %FFP was affected by the choice of algorithm ($p < 0.001$).

*Table 1. Agreement of automatic outlines with acute manual outlines in training and validation groups. * P < 0.05 compared to all other algorithms*

	Training (N=116)			Validation (N=122)		
Classifier	**Sensitivity(%)**	**Specificity(%)**	**DSC**	**Sensitivity(%)**	**Specificity(%)**	**DSC**
ADC thr.	53.5 [10.4-78.5]	99.76 [99.32-99.95]	0.53 [0.10-0.74]	24.9 [0.5-50.2]	100 [99.96-100]	0.37 [0.01-0.64]
k-means	30.2 [0-74.0]	99.94 [99.57-100]	0.24 [0-0.67]	35.7 [0-64.8]	99.96 [99.64-100]*	0.20 [0-0.66]
ISODATA	40.7 [0-76.3]	99.91 [99.57-100]	0.36 [0-0.71]	0 [0-47.9]	100 [99.93-100]	0 [0-0.44]*
k-NN	39.7 [6.8-60.9]	99.98 [99.90-100]	0.50 [0.08-0.73]	15.5 [0-42.6]	100 [99.98-100]	0.27 [0-0.59]
NB	74.6 [53.3-89.6]*	99.28 [98.58-99.75]*	0.58 [0.21-0.79]	58.5 [20.2-78.2]*	99.77 [99.48-99.94]*	0.55 [0.12-0.80]*

Table 2. Proportion of FFP and FFN tissue in validation group

Classifier	**FP(cm^3)**	**FN(cm^3)**	**FFP(%)**	**FFN(%)**
ADC thr.	0.04 [0-0.44]	4.88 [1.24-16.27]	52.8 [5.6-87.5]	54.8 [33.9-75.6]
k-means	0.38 [0.02-3.84]	3.67 [1.03-11.74]	7.9 [0-70.8]	54.8 [37.6-79.1]
ISODATA	0 [0-0.45]	5.04 [1.05-17.19]	21.8 [0.4-75.0]	54.5 [31.2-76.6]
k-NN	0 [0-0.16]	5.05 [1.365-16.33]	33.3 [0-100]	53.8 [36.0-73.1]
NB	1.94 [0.51-4.99]	2.64[0.76- 7.64]	9.3 [0-35.9]	63.6 [38.9-82.1]

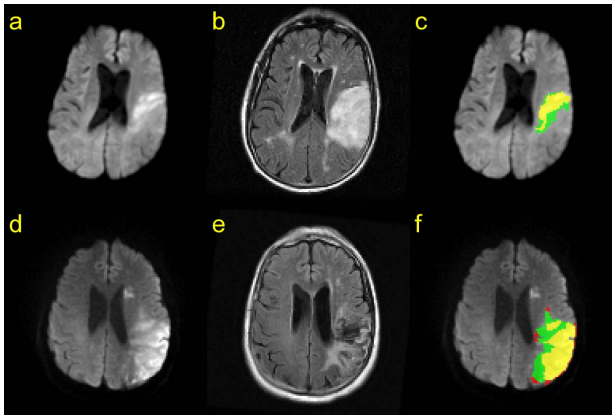

Fig. 1. Top row: a) acute DWI and b) FU FLAIR; c) false positives of k-means compared with FMO with yellow = TP with AMO, green = FFP and red = TFP (none in this slice). Bottom row: d) acute DWI and e) FU FLAIR; c) false negatives of NB compared to FU, with yellow = TP, green = FFP and red = TFP

Discussion: The relative performance of the algorithms compared to AMO agrees with previous work, with NB having greater sensitivity and DSC, but lower specificity, compared to other methods. Sensitivity and DSC of all approaches tended to be somewhat inferior overall compared to the training group, especially for ISODATA and k-NN. However, specificities were higher. In a substantial number of cases, significant fractions of FP and FN voxels appeared to be respectively FFP and FFN when the FU lesion was considered, suggesting our automated approaches may be more accurate than manual outlines.

References: [1] Neumann-Haefelin T et al, Ann Neurol. 2000;47(5):559-70; [2] Schellinger P et al, Neurology, 2010; 75:177-185; [3] Mocking S et al, Proc. ISMRM 2011; [4] Mocking S et al, Automatic segmentation of diffusion MRI in 116 cases of ischemic stroke. In preparation; [5] Sorensen et al, Radiology 1999; 212(3):785-92; [6] Collins DL et al. Comput. Assist. Tomogr. 1994; 18, 192-205. [7] Straka M et al, JMRI, 2010; 32:1024-1037; [8] Ball GH, Storming Media, 1965; [9] Cover TM, Hart PE, IEEE Trans. Inf. Theory 1967; 13(1):21-27. [10] Maron ME, Journal of the ACM, 1961; 8(3):404-417.