Journal of Neuroradiology 46 (2019) 288–293

Available online at ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

Original Article

# Comparison of automated and visual DWI ASPECTS in acute ischemic stroke

E. Kellner[a,*], M. Reisert[a], V.G. Kiselev[a], C.J. Maurer[b], H. Urbach[c], K. Egger[c]

[a] Department of Radiology, Medical Physics, Medical Center, University of Freiburg, Faculty of Medicine, University of Freiburg, Germany
[b] Department of Diagnostic and Interventional Radiology and Neuroradiology, Augsburg Hospital, Germany
[c] Department of Neuroradiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany

ARTICLE INFO

*Article history:*
Available online 9 March 2019

ABSTRACT

*Background and purpose.* – To assess intra-and inter-rater agreement of the ASPECTS (Alberta Stroke Program Early CT Score) based on diffusion-weighted MRI and to compare it with fully – automated methods (eASPECTS).
*Methods.* – DWI-ASPECTS of scans of 96 patients with acute ischemic stroke was rated by 2 experts. Automated methods based on thresholding the affected volumes of a coregistered atlas, and a regression tree learning method were established. Intra-rater, inter-rater and human-rater vs. automated methods agreements were investigated based on the intraclass correlation coefficients (ICC) and Bland Altman plots.
*Results.* – Intra-rater agreement was good for both raters (ICC of 0.91 and 0.93). Inter-rater agreement was worse (ICC = 0.86) indicating a slight bias between both raters. Agreement with automated methods ranged from 0.81 to 0.87. Root-mean-squared deviation was 0.89 and 0.69 for the human raters and ranged from 0.95 to 1.24 for the automated methods.
*Conclusions.* – Agreement values are on the same order or higher compared to a literature review of CT-based ASPECTS. Automated methods perform slightly worse than human expert ratings, but they still have enough power to determine the DWI-ASPECTS with good precision in a clinical setting.
© 2019 Elsevier Masson SAS. All rights reserved.

## Introduction

The choice of an imaging criteria to select patients for mechanical thrombectomy is still a matter of debate [1–3]. Using native CT, the infarct core size is usually approximated by the Alberta Stroke Program Early CT Score (ASPECTS [4]). The ASPECT score was used as an inclusion criterion in three of the major thrombectomy studies conducted in 2015: ESCAPE [5], REVASCAT [6] and SWIFT PRIME [7]. Nevertheless, diffusion-weighted MRI is superior to native CT in detecting ischemia: a Cochrane review published in 2009 gives a sensitivity of 0.99 for DWI (95% confidence interval 0.23 to 1.00) and a sensitivity of 0.39 for CT (95% confidence interval 0.16 to 0.69) [8]. The recently published DAWN study, in which patients benefit from mechanical thrombectomy even between 6 to 24 hours after clinical symptom onset, was based on extended imaging with DWI or CT perfusion demonstrating a small infarct core as essential inclusion criteria, and has once again highlighted the need for reliable patient selection [9,10]. The ASPECTS is a 10 point score that measures the extent of early ischemic changes in the middle cerebral artery territory on CT [11] and is significantly correlated with clinical outcome [12]. However, as ASPECTS scoring on CT is user- and experience-dependent, the inter-rater agreement especially within the first hours after stroke onset is low [13]. Based on a meta-analysis including 30 articles, intra-rater agreement was slight to moderate ($\kappa = 0.042–0.469$). Even when ASPECTS was dichotomized as 0–5 vs. 6–10, inter-rater agreement did not reach a substantial level ($\kappa = 0.561$) [14]. Regarding additional use of automated computer-based ASPECT scoring (eASPECTS) intra-rater agreement remains moderate with intraclass correlation coefficients between eASPECTS and raters of 0.72, 0.74, and 0.76, respectively [12]. Since diffusion-weighted MRI (DWI) is more sensitive for the detection of early ischemic changes than CT, aim of this study was to evaluate ASPECTS reading based on DWI (DWI-ASPECTS) and compare it to automated segmentation- and machine-learning-based results (DWI-eASPECTS).

* Corresponding author at: University Medical Center Freiburg, Department of Radiology, Medical Physics, Breisacherstr. 60a, 79106 Freiburg, Germany.
*E-mail address:* elias.kellner@uniklinik-freiburg.de (E. Kellner).

https://doi.org/10.1016/j.neurad.2019.02.006
0150-9861/© 2019 Elsevier Masson SAS. All rights reserved.

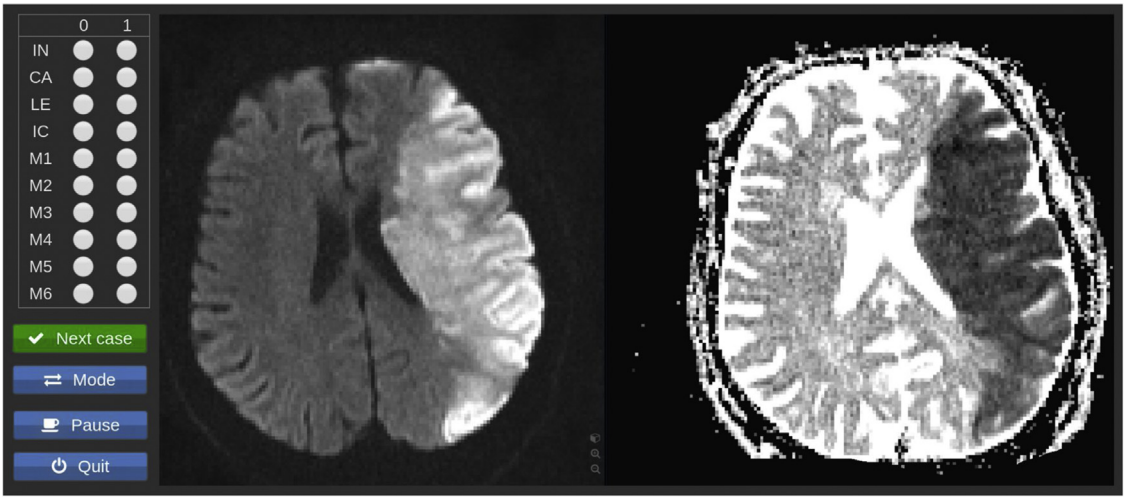


**Fig. 1.** Design of the reading form using NORA (www.nora-imaging.com): DWI and ADC maps were presented, the raters could synchronously scroll through the slices and adjust the contrast. For each region, they had to assess whether it was affected or not. All cases were presented in a randomized order. The reading was repeated twice, with an interval of at least one week.

## Material and methods

### *Patients*

96 patients with acute ischemic stroke and MR angiography-proven M1 segment occlusions were studied on two different MRI scanners (Magnetom Avanto; Magnetom Trio, Siemens, Erlangen, Germany). The study was approved by the local ethics committee. Informed consent was obtained from all individual participants included in the study.

### *MRI*

The MRI stroke protocol included a diffusion-weighted scan (DWI) and a perfusion weighted dynamic-susceptibility-contrast scan (PWI). Typical DWI parameters were: Axial DWI sequence, TR/TE 4900 – 5300 ms/100 ms, *b*-values 0 and 1000s/mm$^2$, resolution $1 \times 1 \times 5$ mm, matrix size $192 \times 192$). Typical PWI parameters were: Injection of contrast agent (0.5 mol/L Gd-DTPA, Multihance, Bracco Imaging, Italy) at a normal dose of 0.1 mmol per kg body weight, followed by 30 mL saline flush, both at a rate of 5 mL/s. EPI time series with TR = 1800 ms, TE = 45 ms, 6/8 partial fourier, matrix size $112 \times 112$, resolution $2 \times 2 \times 5$ mm, 17 slices.

### *Visual rating of the ASPECTS*

Visual rating was performed using the NORA medical imaging platform (www.nora-imaging.com) as follows: The ADC map, the diffusion-weighted image, and a rating form including check-boxes for the ten regions were shown for each case (Fig. 1). All cases were rated by two experienced neuroradiologists (CJM, HU) in a randomized order. Both readers performed a second reading run after an interval of at least one week. An institutional review board approval was obtained for this prospective study.

### *Automated rating of the eASPECTS*

The automated eASPECTS was calculated using the following procedure:

- automated segmentation of the infarction core;
- non-rigid mapping of the ASPECTS atlas to the individual image space;
- calculating the affected volume fractions for each region;
- calculating the final score in two different ways: by thresholding, and by a machine learning based method.
- segmentation of the infarction core was based on the ADC and TMAX maps and performed using the fully automated MRI-Stroke-evaluation pipeline recently published [15]. In brief, the method is based on the commonly used thresholding of the ADC map with ADC < $600 \times 10^{-6}$ mm$^2$/s together with a reduction of the number of false positives by restriction to areas showing a significantly higher TMAX value in the affected hemisphere only.
- the images with $b = 0$ s/mm$^2$ were segmented using SPM 8 (Wellcome Department of Imaging Neuroscience, University College London, UK). The inverse deformation field was then used to map a custom-build atlas to the individual patient space. The atlas is based on the ASPECTS-relevant regions [4]. The atlas as depicted in Fig. 2 was created in MNI space by an experienced neuroradiologists (KE) not participating in the rating.
- the affected volume fractions were calculated in the individual space by overlaying each region with the segmented infarction core, and counting its relative overlap (intersection divided by region volume).
- the 10 affected volume fractions were used as features for final calculation of the DWI- eASPECTS. Here, we followed two approaches: A rather simple thresholding based, and a machine learning based one.

### *Thresholding-based approach*

For the thresholding method, we used a single threshold for all regions, e.g. 10% and additionally, we performed an analysis using two different thresholds, one for [CA, LE, IC, IN] and a different one for [M1, M2, M3, M4, M5, M6] since raters might tend to intuitively apply a different threshold for smaller and larger regions. The thresholds were optimised by minimising the root-mean-squared-deviation (RMSD) as compared to the manual ratings of the 2 neuroradiologists. This was done for both raters independently to account for potential systematic differences between them. For each rater, we used the average of their two runs.

### *Machine-learning-based approach*

For machine learning, we aimed at predicting the total ASPECT score, taking the affected volume fractions of the 10 regions as input

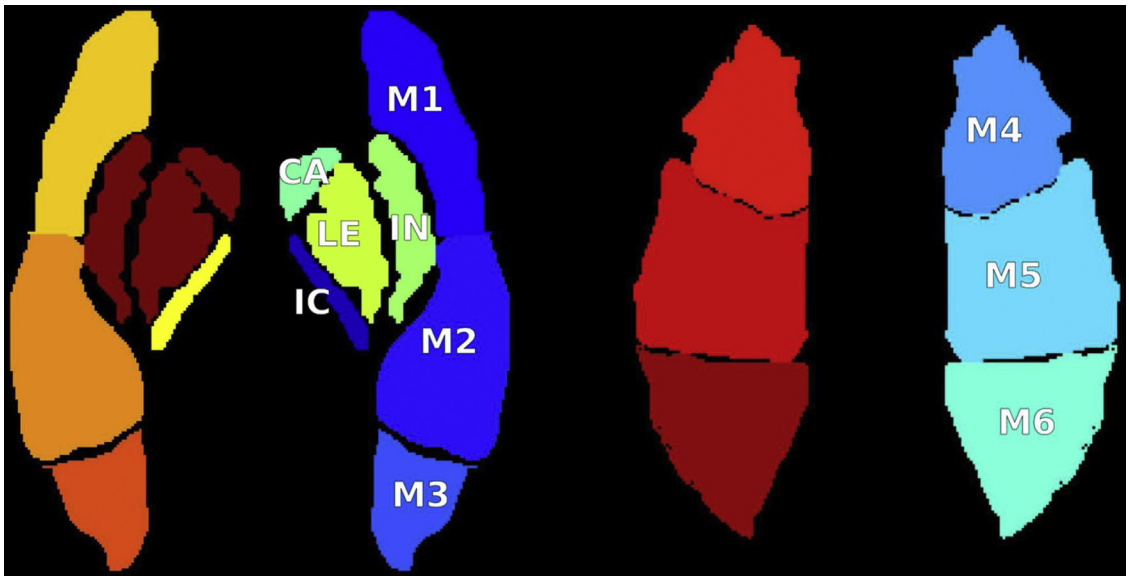


**Fig. 2.** ASPECTS 3D atlas used for the automated calculation of the score. IC: interne capsule; IN: insular ribbon; LE: Lentiform nucleus; CA: caudate nucleus; M1-M6: cortical and subcortical regions covering the territory of the middle cerebral artery. The atlas was created by an experienced neuroradiologist not participating as a rater. The 3D atlas was mapped to the individual spaces for each case, and the region-wise overlap with the segmented infarction core was calculated to get the affected volume fraction for each region.

features. We focused on decision tree learning, which is a powerful, but yet relatively simple approach. Since the ASPECTS is defined on an interval scale, we chose to use a regression tree. The concept behind this technique is to build a tree where nodes correspond to decisions which are based on testing for certain conditions, for example, volume fraction in IC > 0.34 and volume fraction in M1 < 0.63 and ... The leaves correspond to predictions. Learning comprises building the tree, i.e. defining the node structure and setting certain values for the thresholds for each node. There are various concepts and strategies to obtain a stable solution, such as tree-bagging and boosting [16]. We tested several of these strategies and also investigated the N-1-fold loss curves. As a result, we found that best performance was obtained with a boosting strategy based on least-squares ('LSBoost'), starting with an ensemble of 100 trees. Here, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown already. The ensemble fits to minimize the mean-squared error. For references related to LSBoost, see [17], Chapters 7 (Model Assessment and Selection) and 15 (Random Forests, see also [18]). For the statistical analysis of the prediction errors, we applied N-1-fold cross-validation ("leave-one-out"). Technically, we used an out-of-the-box solution called "fitensemble", provided by the MATLAB statistics toolbox (MATHWORKS, Natick, USA). The learning was conducted for both readers independently, based on the average of their two runs.

### *Statistical analysis*

Statistical analysis aimed at comparing intra-rater reproducibility for the repeated runs, the inter-rater agreement between both raters and the agreement between the human raters and the eASPECTS. For that, we calculated three different agreement measures: The root-mean-squared deviation (RMSD), Pearson's correlation coefficient (r), and the intra-class-correlations coefficient (ICC) type 2-1, which also accounts for absolute differences between two ratings. Finally, we calculated Fleiss Kappa statistics to compare our results with the meta-analysis by Farzin et al. [14]. For visualization of the agreements, we created correlation plots and Bland–Altman plots.

## Results

All results are summarized in Table 1 and in the correlation- and Bland-Altman plots, Fig. 3 and Fig. 4.

Human Raters: The reproducibility for rater A was: $r = 0.91$, ICC = 0.91 RMSD = 0.89, for rater B, it was $r = 0.93$, ICC = 0.93 RMSD = 0.69. Correlation between the raters was = 0.90, ICC = 0.86 RMSD = 1.02. The difference between r and the ICC indicates an overall small systematic difference between the two readers. This can also visually be observed at correlation plot A vs. B in Fig. 3. The results for comparison with the eASPECTS were as follows: Threshold Based: Rater A, one threshold, optimal threshold 7%: correlation parameters $r = 0.87$, ICC = 0.86 RMSD = 1.16. Rater B, one threshold, optimal threshold 10%: $r = 0.85$, ICC = 0.82 RMSD = 1.24. Rater A, two thresholds, Optimal values 8% and 6.5%: $r = 0.88$, ICC = 0.87 RMSD = 1.13. Rater B, two thresholds, Optimal values 42% and 4.5%: $r = 0.83$, ICC = 0.81 RMSD = 1.15. Regression Tree: Rater A: $r = 0.84$, ICC = 0.83 RMSD = 1.08. Rater B: r = 0.85, ICC = 0.84 RMSD = 0.95.

## Discussion

The goal of this study was

- to assess the reproducibility and inter-rater correlation of a visual DWI - ASPECTS rating;
- to compare it with automated methods (DWI – eASPECTS).

The intra-rater agreement on the order of r = 0.92, ICC = 0.92 suggests that the DWI-ASPECTS can be used as a robust decision criterion in terms of reproducibility. The agreement between the raters, $r = 0.90$ and ICC = 0.86, is slightly worse, with the ICC being lower than Pearson's r indicating a slight systematic difference between the two raters, as the ICC accounts also for absolute differences. This systematic bias emphasizes that if ASPECTS is to be used as a standardised quantity between different centers and raters, a proper definition of the ASPECTS regions together with a corresponding training of radiologists is necessary.

The optimal threshold for the eASPECTS using a single threshold is 7% for rater A, and 10% for rater B. These values seem to be on a reasonable order. (The different values again show the slight sys-

**Table 1**
Results of all correlation measures. Obviously, rater B has the best reproducibility with ICC = 0.93 and RMSD = 0.69. Compared to the Inter-rater agreement A vs. B, all automated methods perform slightly worse, with the machine learning approach showing the best performance.

| | A vs. A | B vs. B | A vs. B | A vs. E np = 1 | B vs. E np = 1 | A vs. E np = 2 | B vs. E np = 2 | A vs. E ML | B vs. E ML |
|---|---|---|---|---|---|---|---|---|---|
| r | 0.91 | 0.93 | 0.90 | 0.87 | 0.85 | 0.88 | 0.83 | 0.84 | 0.85 |
| ICC | 0.91 | 0.93 | 0.86 | 0.86 | 0.82 | 0.87 | 0.81 | 0.83 | 0.84 |
| RMSD | 0.89 | 0.69 | 1.02 | 1.16 | 1.24 | 1.13 | 1.15 | 1.08 | 0.95 |
| Fleiss kappa | 0.47 | 0.53 | 0.33 | 0.27 | 0.27 | 0.24 | 0.21 | 0.35 | 0.31 |

A: Rater 1; B: Rater 2; E: eASPECTS; np: 1: one threshold; np: 2: 2 thresholds; ML: machine learning; r: Pearson's r; ICC: intra-class-correlation-coefficient type 2-1; RMSD: root-mean-squared-deviation.

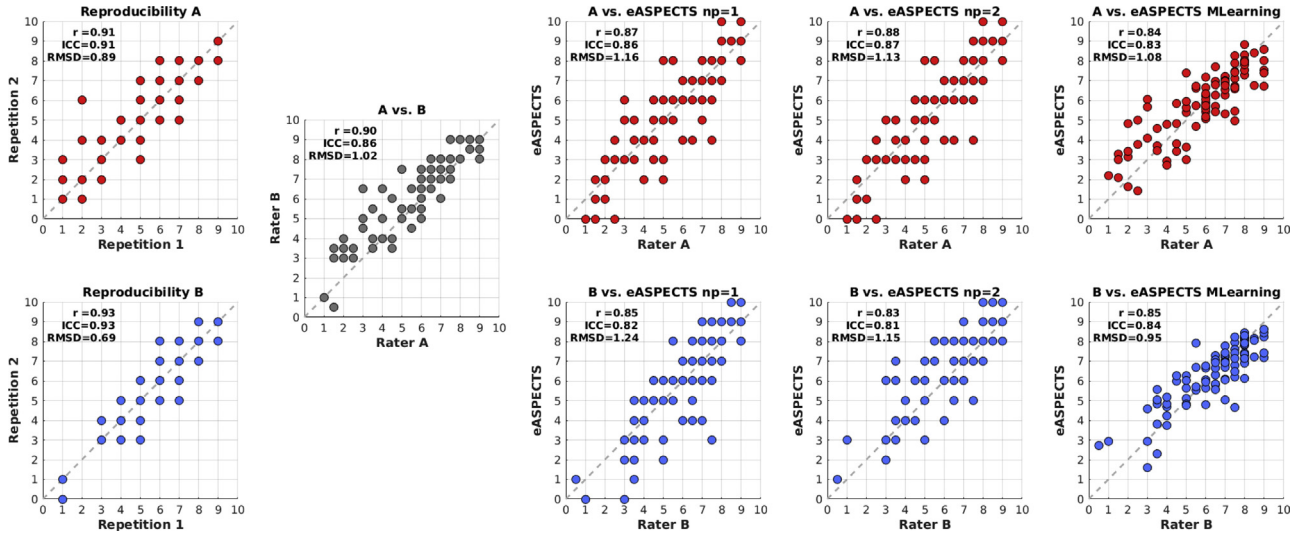


**Fig. 3.** Correlation plots. Rater A and B (top and bottom row) are highlighted in different colors. Obviously, rater B has the best reproducibility with RMSD = 0.69. Compared to the Inter-rater agreement A vs. B, all automated methods perform only slightly worse, with the machine learning approach showing the best performance. Note that ASPECTS values are discrete (except for Machine Learning), so one point can correspond to multiple cases, especially in the very frequent range around ASPECTS = 7. Only the machine-learning approach predicts continuous values, such that the distribution of values becomes more obvious there. A. Rater 1; B. Rater 2; E. eASPECTS; np = 1: one threshold; np = 2: 2 thresholds; MLearning. machine learning; r. Pearson's r; ICC. Intra-class-correlation-coefficient type 2-1; RMSD. root-mean-squared-deviation.

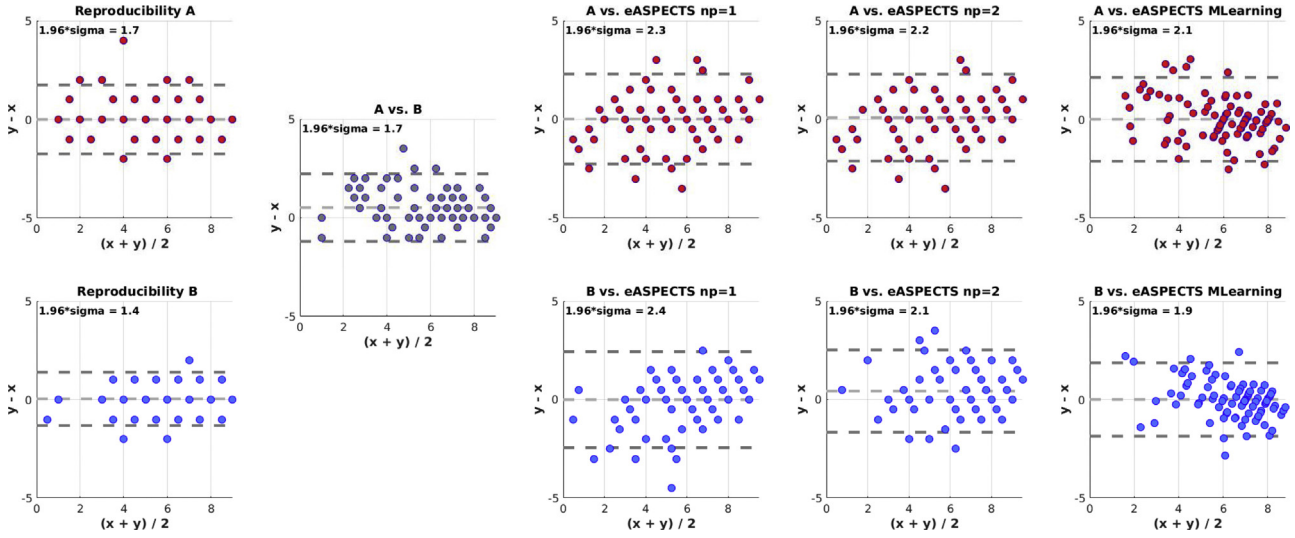


**Fig. 4.** Bland–Altman plots of the representation in Fig. 3. As in Fig. 3, Rater A and B (top and bottom row) are highlighted in different colors. Sigma corresponds to the RMSD. The interval range [−1.96 sigma, 1.96 sigma] indicates the confidence interval. Obviously, as in Fig. 3, rater B has the best reproducibility indicated by the smallest RMSD. Compared to the Inter-rater agreement A vs. B, all automated methods perform only slightly worse, with the machine learning approach showing the best performance. Note that ASPECTS values are discrete (except for Machine Learning), so one point can correspond to multiple cases, especially in the very frequent range around ASPECTS = 7. A: Rater 1; B: Rater 2; E: eASPECTS; np: 1: one threshold; np: 2: 2 thresholds; MLearning: machine learning; sigma: standard deviation of X-Y.

tematic difference between the two raters: Rater B tends to count ASPECTS points more "stringent" than rater A.) With two thresholds, one for the larger and one for the smaller regions, the precision is slightly increased. The optimal thresholds can be understood in the sense that Rater A counts smaller regions from a value 8% and larger regions from 6.5%, whereas rater B counts smaller regions from 42%, and the larger ones from 4.5%. The great value of 42% for rater B already gives a hint towards the ill-posedness of the prob-

lem with a flat minimization landscape: The optimal thresholds in this case might be good to explain the data, but are not necessarily plausible in their interpretation as volume fractions. Therefore, a model using more thresholds - 10 in the most extreme case - might further improve the fitting, but is hard to be generalized in terms of suggesting a simple optimal value with a clear meaning. Keeping this in mind, it seems that with the plausible and simple single value of 7%–10% percent for any region, one can already obtain quite reasonable results.

The automated DWI-eASPECTS using a machine learning approach based on a regression tree resulted in an improved root-mean-squared-deviation (RMSD) of 1.08 for rater A, and 0.95 for rater B compared to the single threshold-based approach (1.16 and 1.24, respectively), which indicates an even more accurate prediction of the ASPECTS. However, machine-learning based methods come at the cost of a less-straightforward interpretability and it is more difficult to reproduce them in other studies and sites. In view of the only minor improvement of the machine learning based approach, from a clinical and practical perspective, we suggest that the simple single-threshold based method is already sufficient to predict the eASPECTS with a reasonable accuracy in a clinical setting and has the advantage of a clear interpretability and standardizability.

All automated methods were based on using the affected volume fractions as input features. These fractions were obtained by an automated segmentation algorithm to delineate the infarction core and overlay it with an ASPECTS atlas. Hence, any error during segmentation or co-registration with the atlas will propagate to the final score. In principle, these problems might be overcome by a more generalized learning approach operating directly on the full images, instead of segmentations, and an atlas. Recent developments in deep-learning have shown a great potential in this regard and might be used together with a larger training data set to further improve the automated methods in future works.

Concerning the comparison with other studies, the meta-analysis by Farzin et al. [14] can serve as a good benchmark. They analysed the intra- and inter-rater agreement of the ASPECTS based on the data of as much as 30 articles. They report intra-rater ICCs ranging from 0.59 to nd 0.94 (mean 0.81), and inter-rater ICCs between 0.672 and 0.811 (mean 0.68). With respect to this, our ICC values (intra-rater mean 0.92 and inter-rater 0.86) are on the upper edge and higher than the mean values. This again indicates that DWI ASPECTS ist more specific, as most of the articles investigated by Farzin et al. deal with native CT images. The Fleiss kappa values show the same pattern: Our values of 0.5 for the intra-rater, and 0.33 for the inter-rater agreement are higher than values from the meta-study ranging from 0.04–0.47 for the intra-rater, and 0.13–0.33 (mean 0.18) for the inter-rater agreement.

It should be noted that a limiting factor of the present study is that visual rating was performed by two raters only, whereas the number of raters ranged from 2 to 15 in the studies listed by Farzin. Increasing the number of raters would allow for a more precise estimate of a potential bias between raters, which is linked to the issue of a proper, universal definition of the ASPECTS, as discussed above. In a recent work similar to the present study, Schröder et al. [19] presented a very interesting approach based on a probabilistic template to achieve the goal of a more universal and data-driven definition of an ASPECTS atlas. They report optimal relative lesion loads (RLL) of $7.7 \pm 12.6\%$ for automated ASPECTS counting. This seems to be in very good agreement with the optimal single-threshold values from our study (7% and 10% for rater A and B). Further, their finding that the RLL of subcortical regions (C, IC, I, and L) was significantly higher than for the cortical regions (M1–M6) is in line with our results from the 2-threshold approach, which suggested a higher threshold for subcortical regions, even though this result is not very decisive, as discussed above. Finally, Schröder et al. report kappa-values of around 0.39 for a threshold of 10%, which is close to the kappa-values we obtained in our study.

Farzin et al. are rather sceptical considering reliability and the usefulness of the ASPECTS in general. However, we do not fully agree with this interpretation. Considering for example the RMSD – i.e. the standard deviation in the correlation plots – in our results, we interpret that the DWI-ASPECTS can roughly be estimated up to plus/minus one ASPECTS point, with the human ratings presumably having a slightly better precision than the automated methods. We conclude that the DWI-ASPECTS has a greater precision and reproducibility than the original, CT based ASPECTS, and that it can automatically be determined in a robust manner.

## Conclusions

In this work, we evaluated a DWI based ASPECTS reading study with two raters and compared it with an automatically derived DWI-eASPECTS. The results indicate that reproducibility and inter-rater agreement seems good enough to use the DWI ASPECTS in clinical routine for decision support. The automated scoring showed almost comparable results to human ratings and might hence be used as a standardised and automated marker for decision support in acute stroke.

## Disclosure of interest

The authors declare that they have no competing interest.

## Acknowledgements

We are grateful to Hansjörg Mast for his support with the measurements.

## References

[1] Powers WJ, Derdeyn CP, Biller J, Coffey C, Hoh BL, Jauch EC, et al. 2015 American Heart Association/American Stroke Association focused update of the 2013 guidelines for the early management of patients with acute ischemic stroke regarding endovascular treatment. Stroke 2015;46(10):3020–35.
[2] Wahlgren N, Moreira T, Michel P, Steiner T, Jansen O, Cognard C, et al. Mechanical thrombectomy in acute ischemic stroke: Consensus statement by ESO-Karolinska Stroke Update 2014/2015, supported by ESO, ESMINT, ESNR and EA. Int J Stroke 2016;11(1):134–47.
[3] Wintermark M, Luby M, Bornstein NM, Demchuk A, Fiehler J, Kudo K, et al. International survey of acute stroke imaging used to make revascularization treatment decisions. Int J Stroke 2015;10(5):759–62.
[4] Pexman JH, Barber PA, Hill MD, Sevick RJ, Demchuk AM, Hudon ME, et al. Use of the Alberta Stroke Program Early CT Score (ASPECTS) for assessing CT scans in patients with acute stroke. AJNR Am J Neuroradiol 2001;22(8):1534–42.
[5] Goyal M, Andrew MD, Menon BK, Eesa M, Rempel J, Thornton J, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. N Engl J Med 2015;372(11):1019–30.
[6] Jovin TG, Chamorro A, Cobo E, de Miquel MA, Molina CA, Rovira A, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. N Engl J Med 2015;372(24):2296–306.
[7] Saver JL, Goyal M, Bonafe A, Diener HC, Levy EI, Pereira VM, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. N Engl J Med 2015;372(24):2285–95.
[8] Brazzelli M, Sandercock PA, Chappell FM, Celani MG, Righetti E, Arestis N, et al. Magnetic resonance imaging versus computed tomography for detection of acute vascular lesions in patients presenting with stroke symptoms. Cochrane Database Syst Rev 2009.
[9] Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. N Engl J Med 2017;378(1):11–21.
[10] Jovin TG, Saver JL, Ribo M, Perreira V, Furlan A, Bonafe A, et al. Diffusion-weighted imaging or computerized tomography perfusion assessment with clinical mismatch in the triage of wake up and late presenting strokes undergoing neurointervention with Trevo (DAWN) trial methods. Int J Stroke 2017;12(6):641–52.
[11] Barber PA, Demchuk AM, Zhang J, Buchan AM, ASPECTS Study Group. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. The Lancet 2000;355(9216):1670–4.

[12] Pfaff J, Herweh C, Schieber S, Schönenberger S, Bösel J, Ringleb PA, et al. e-ASPECTS correlates with and is predictive of outcome after mechanical thrombectomy. AJNR Am J Neuroradiol 2017, http://dx.doi.org/10.3174/ajnr.A5236.
[13] Bal S, Bhatia R, Menon BK, Shobha N, Puetz V, Dzialowski I, et al. Time dependence of reliability of noncontrast computed tomography in comparison to computed tomography angiography source image in acute ischemic stroke. Int J Stroke 2015;10(1):55–60.
[14] Farzin B, Fahed R, Guilbert F, Poppe FY, Daneault N, Durocher AP, et al. Early CT changes in patients admitted for thrombectomy Intrarater and inter-rater agreement. Neurology 2016;87(3):249–56.
[15] Kellner E, Reisert M, Kiselev VG, Maurer CJ, Beume L, Urbach H, et al. Automated infarct core volumetry within the hypoperfused tissue: technical implementation and evaluation. J Comput Assist Tomogr 2017;41(4):515–20.
[16] Bauer E, Ron K. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 1999;36(1):105–39.
[17] Breiman L. Random forests. Machine Learning 2001;45(1):5–32.
[18] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. second edition New York: Springer; 2008.
[19] Schröder J, Cheng B, Malherbe C, et al. Impact of lesion load thresholds on Alberta stroke program early computed tomographic score in diffusion-weighted imaging. Front Neurol 2018;9:273 [https://doi.org/10.3389%2Ffneur.2018.00273].