More is better: Rater consensus on lesion outline in acute stroke improves predictive models

K. Y. Jonsdottir[1], L. Østergaard[1], and K. Mouridsen[1]
[1]Department of Neuroradiology, Aarhus University Hospital, Aarhus, Aarhus C, Denmark

Introduction

Correct delineation of the final infarct resulting from acute ischemic stroke on T2 FLAIR images is essential in clinical studies using lesion volume or location as endpoints. In particular, voxel-based models predicting the final infarct based on acute MRI parameters may depend critically on correct outcome classification of the training data. While studies typically engage only a single rater for lesion outlining, we hypothesize that performance of predictive algorithms increases when final outcome is determined as the consensus between multiple raters. Here we demonstrate increments in several performance metrics by requiring consensus between up to nine expert neuroradiologists.

Materials and methods

Final infarcts were delineated on follow-up (FU) FLAIR images by 9 raters in 14 patients. Nine surrogate FU masks were generated such that all voxels in FU_n were labelled/considered infarcted by at least n raters, n =1,…,9 (cf. Figure 1). The measured lesion volume (MLV) of the FU masks were compared. Nine logistic regression models [1,2] were developed using acute structural, DWI and PWI images and FU_1 ,..., FU_9 as end lesion, representing the different degrees of consensus. For each model the predicted lesion volume (PLV) was calculated at a 50% threshold. The PLVs were compared to the MLVs of each FU_n, n=1,…,9. Performance in predicting the FU masks was quantified using three performance measures: area under the ROC curve (AUC), a prediction-index (p-index) measuring the difference in mean infarction risk of infracting and surviving voxels at risk and finally the variance of the p-index (v-index). High values of AUC and the p-index in combination with a low v-index indicate effective (accurate) predictions of the final infarct. Performance measures are calculated using ipsilateral tissue affected by the stroke. All pair-wise comparisons are done using a signed rank test.

Results

The volume of FU_1 was in median 260% larger than the volume of the most conservative lesion mask FU_9, illustrating large volume variability in individual outlines. For each model the correlation between PLVs and MLVs was highest for FU_9 and lowest for FU_1. The highest correlation was 0.86 (p=0.0001) between the MLV of FU_9 and the PLV applying model using a complete consensus and the lowest was 0.55 (p=0.04) between the MLV of FU_1 and PLV applying model using minimal consensus. Figure 2a shows that AUC increases with the degree of consensus in the FU mask, with a maximum of 0.82 predicting the most conservative lesion volume FU_9. For each model, the increase was significant ($p<0.05$) for degrees of consensus between 1 and 5. Similarly, Figure 2b shows the p-index also increases with the degree of consensus in the FU mask. Moreover, given an arbitrary FU mask, the p-index is largest for the model trained using the most conservative lesion mask. The increase is significant ($p<0.05$) in both directions for most degrees of consensus. Finally, the v-index shown in Figure 3c is quite small. It varies from $3x10^{-5}$ to 10^{-4} taking maximum values where the p-index is highest.

Figure 1

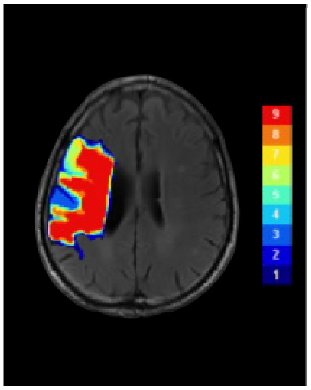
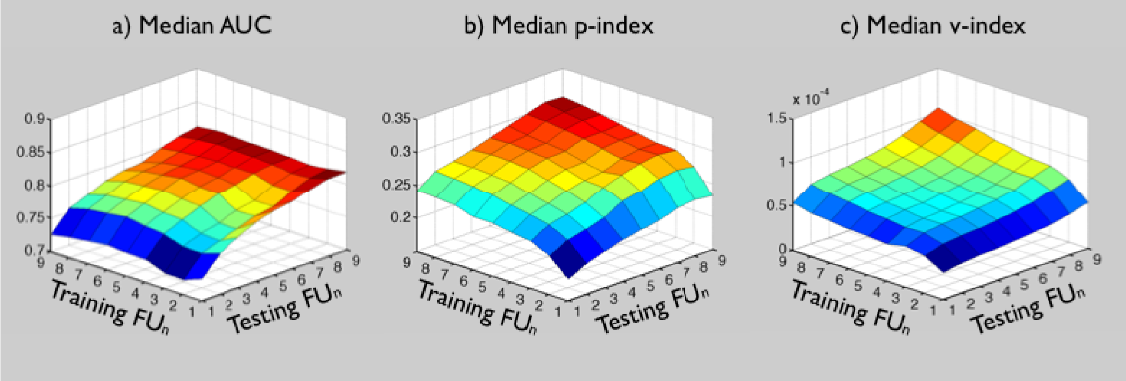


Figure 2



Conclusion

Predictive algorithms depend critically on the manual delineation of the final infarct in training data. Performance is optimized if the FU lesion is defined based on a complete rater agreement, in which case the relation between acute parameters and final outcome seemingly is better identified by prognostic models. Consequently, we suggest employing multiple raters or conservative lesion definitions in devising predictive algorithms. As multiple rating is typically time consuming and impractical, construction of automatic lesion detection algorithms facilitating comparisons between studies, should be pursued. Alternatively, standard guidelines for outlining of the final lesion could be established.

References

[1] Wu et al. *Stroke* (2001). 32:933-42. [2] Wu et al. *Brain* (2006). 129:2384-2393.