Multi-scale feature reduction and semi-supervised learning for parsing neuroanatomical heterogeneity

Junhao Wen[1,2], Ganesh Chand[1,2], Ahmed Abdulkadir[1,2], Raymond Pomponio[1,2], Guray Erus[1,2], Aristeidis Sotiras[1,2], Dhivya Srinivasan[1,2], Jimit Doshi[1,2], Alessandro Pigoni, Paola Dazzan, Rene S. Kahn, Hugo G. Schnack, Marcus V. Zanetti, Eva Meisenzahl, Geraldo F. Busatto, Benedicto Crespo-Facorro, Christos Pantelis, Stephen J. Wood, Chuanjun Zhuo, Russell T. Shinohara[2], Haochang Shou[2], Yong Fan[1,2], Ruben C. Gur[1], Raquel E. Gur[1], Dominic B. Dwyer, Theodore D. Satterthwaite, Nikolaos Koutsouleris, Daniel H. Wolf, Erdem Varol[1,2] and Christos Davatzikos[1,2]

1. Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
2. Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

junhao.wen@pennmedicine.upenn.edu

Group comparisons have identified many neuroimaging biomarkers of various diseases, over the past decades, including schizophrenia (SCZ) [1, 2]. However, much of the non-random inter-individual heterogeneity within diagnostic groups is concealed by those analyses which only look at differences across groups with the assumption of groups being homogeneous. Individualized patient classification, commonly performed via machine learning methods, is also challenged by disease heterogeneity, especially if interpretable patterns are sought. The current study aims to disentangle neuroanatomical heterogeneity by leveraging a recently developed semi-supervised machine learning method called HYDRA [3] along with principal component analysis (PCA), seeking the best set of features that enable semi-supervised clustering to detect robust disease subtypes.

• Methods

T1-weighted images from three sites of the PHENOM (Psychosis Heterogeneity via dimEnsional NeurOiMaging) consortium [controls = 364 (age = 29.5±7.0 years; 44.2% female); SCZ = 307 (age = 30.9±7.3 years; 35.2% female)] were analyzed. Brain tissue segmentation was performed using a multi-atlas segmentation technique [4], and were then transformed to produce tissue density maps for volumetric pattern analysis. Gray matter (GM) tissue density maps were harmonized by estimating site, age and gender effects in a pooled sample of matched controls using a voxel-wise linear model.

PCA was applied to the GM maps of the whole population with varying number of components (C = 2 to 40). The extracted components were fed to HYDRA with varying number of clusters (K= 2 to 8). A repeated holdout cross-validation (CV) with 100 repetitions was performed. For each repetition, 80% of the data was used for training HYDRA. During each repetition, the patients of the remaining data (20%) were ignored by HYDRA for consensus clustering assignment [3], thus guaranteeing absence of feature selection bias [5] (The python code will be made publicly available at the time of the conference at https://github.com/anbai106/pyHYDRA). The clustering stability was evaluated using adjusted Rand index (ARI). Statistical mapping was performed using MIDAS [6] to identify the neuroanatomical differences between controls and subtypes in a voxel-wise fashion. False discovery rate correction (FDR) was performed to adjust for multiple comparisons with a significance level of 0.05. Effect size (Cohen's d) was used to visualize the neuroanatomical patterns for those voxels survived at the FDR correction.

• Results

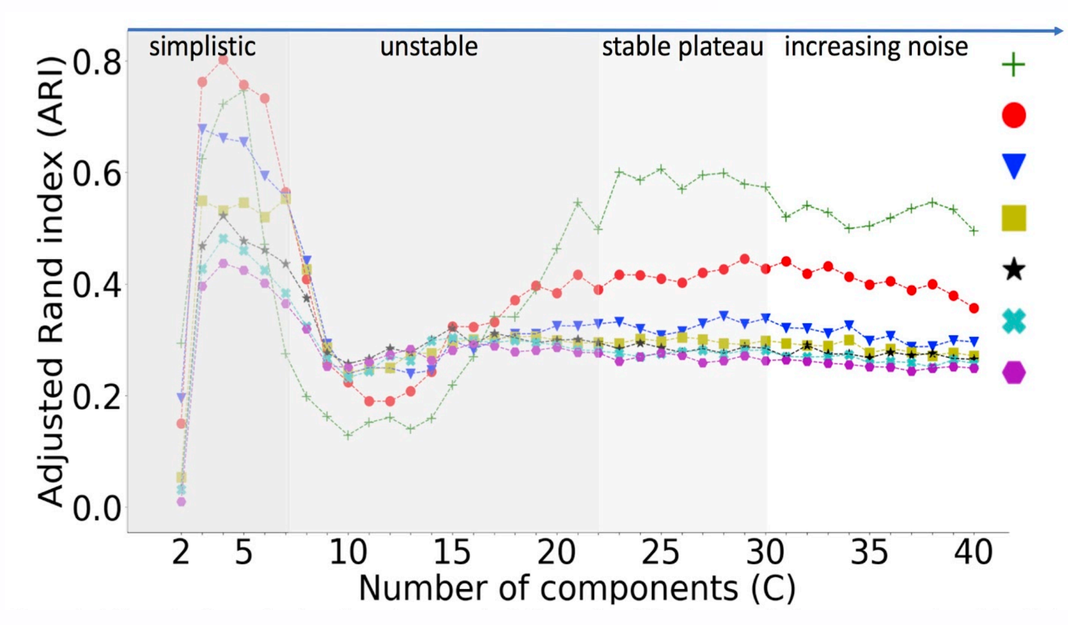

Figure 1. ARI results for evaluating clustering reproducibility and stability. Low-resolution representations (simplistic phase) resulted in high ARIs but the patterns were meanigless and associated with site, age and sex. Unstable phase gave mixed ARI results across different Ks. In consolidation phase, K=2 obtained consensus higher ARI compared to other resolutions. The plateau finally arrived at saturation phase.

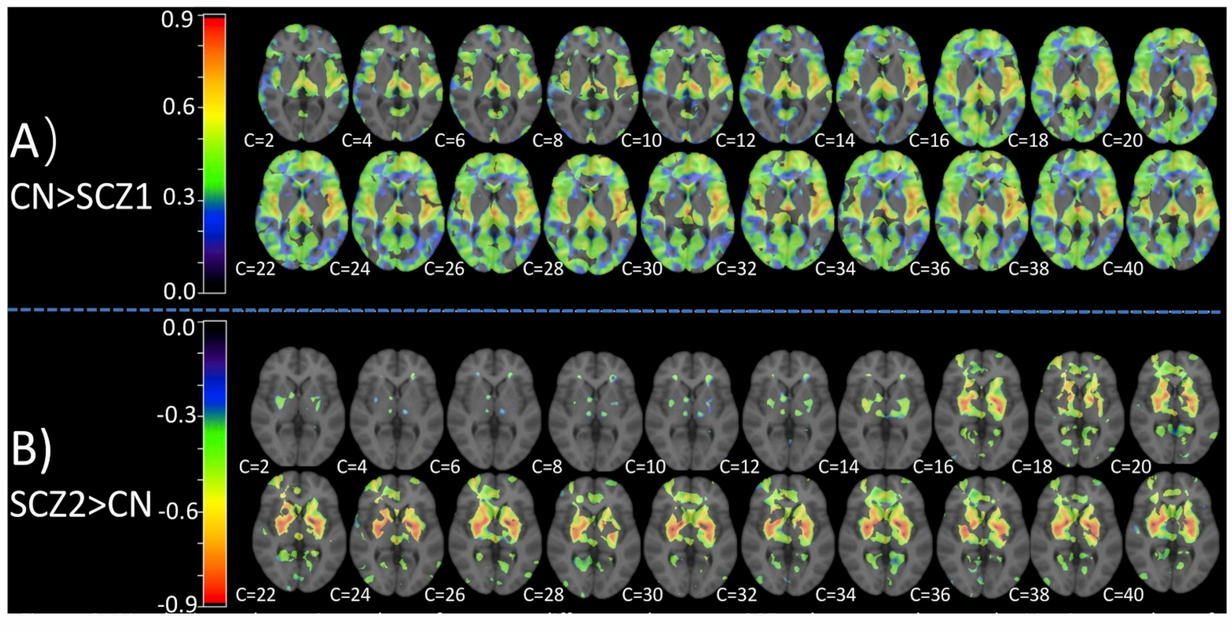

Figure 2. Voxel-wise volumetric analyses for group difference between SCZ subtypes and controls. Varying number of components (C=2 to 40 with step size 2) were displayed. A) CN vs subtype 1, thalamus, insula and diffuse cortical atrophy were found; B) CN vs subtype 2, Larger volumes were found in basal ganglia in subtype 2 patients.

We found two subtypes of SCZ with high clustering reproducibility and stability. Higher ARI was obtained for 2 clusters (K=2) compared to K=3 through 8, for mid to high resolution (C > 19). Coarse resolutions (C between 2, 7) were unable to produce meaningful results (Figure A). In group analysis, Subtype 1 showed widespread GM atrophy, including thalamus, insula and diffuse cortical regions (FDR-p < 0.05) (Figure 2A). Conversely, Subtype 2 demonstrated larger volumes in basal ganglia (FDR-p < 0.05) (Figure 2B). Moreover, with increased C, the neuroanatomical patterns expanded in an additive way, eventually reaching plateau, at a higher resolution (Figure 2). not detailed and thus may have been done by successive evaluations on the test set.

• Conclusion

Our results demonstrate the use of PCA and HYDRA reveals neuroanatomical heterogeneity at appropriate resolution levels, which is estimated in a data-driven fashion. The application to SCZ revealed two highly distinct neuroanatomical profiles, further demonstrating the importance of disentangling this heterogeneity and more generally of establishing distinct neuroanatomical dimensions in brain diseases.

• References

1. Rodrigues-Amorim, D., (2017), *Journal of psychiatric research*, *93*, pp.37-49.
2. van Erp, T. G., et al. (2016), Molecular Psychiatry 21(4), pp.547-553
3. Varol, E., (2017), *Neuroimage*, *145*, pp.346-364.
4. Doshi, J., (2016), *Neuroimage*, *127*, pp.186-195.
5. Wen, J., (2019), AAIC.
6. Varol, E., (2018), *NeuroImage*, *174*, pp.111-126.