Generalization failure of RSFC-based behavioral prediction in non-European pediatric population

Jingwei Li[1,2,3], Danilo Bzdok[4,5], Jianzhong Chen[3], Angela Tam[3], Leon Qi Rong Ooi[3], Avram Holmes[6,7], Tian Ge[7,8,9], Kaustubh R. Patil[1,2], Simon B. Eickhoff[1,2], B.T. Thomas Yeo[5,10,*], Sarah Genon[1,2,*]

[1]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany [2]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Germany [3]ECE, CSC, TMR, N.1 & WISDM, National University of Singapore, Singapore [4]McGill University, Canada [5]Mila - Quebec Artificial Intelligence Institute, Canada [6]Yale University, USA [7]Massachusetts General Hospital, USA [8]Broad Institute of MIT and Harvard, USA [9]Harvard Medical School, USA [10]Integrative Sciences and Engineering Programme (ISEP), National University of Singapore

#1392

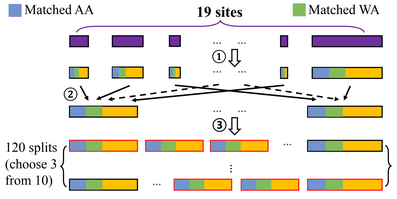
Introduction

Machine learning is expected to play a crucial role in precision medicine, yet algorithmic biases towards majority populations may pose a key challenge towards this goal (Chouldechova 2018; Martin 2019; Obermeyer 2019). In the neuroimaging community, one major line of research is to predict individual's phenotypes from resting-state functional connectivity (RSFC; Finn 2015; Kong 2019; Wu 2020). In that context, predictive models are typically built by capitalizing on large cohorts in which the proportions of certain ethnical groups, e.g. African Americans (AA), are limited.

To evaluate cross-population generalizability of the current, field-standard approach in pre-adolescent populations, we here compared predictive models of behavioral data between phenotypically matched AA and white American (WA) samples in the Adolescent Brain Cognitive Development (ABCD) dataset. When predictive models were trained on the entire sample, out-of-sample prediction errors were generally higher in AA than WA. This bias towards WA corresponds to more WA-like brain-behavioral association patterns learned by models. When models were trained on AA only, compared to training only on WA or an equal number of AA and WA participants, AA prediction accuracy improved but stayed below that for WA.

Methods

- **Dataset:** Adolescent Brain Cognitive Development (ABCD; Volkow 2018; Garavan 2018) (Age: 9-11y, N = 5351, incl. 635 African Americans, 2999 White Americans, 36 behaviors)
- **RSFC preprocessing (Chen 2020):**
  - RSFC across 400 cortical (Schaefer 2018) & 19 subcortical (Fischl 2002) ROIs.
- AA & WA were matched for age, gender, FD, DVARS, intracranial volume (ICV), parental education & behavioral scores.
  - Number of matched pairs of AA and WA ranged from 192 to 301 across behaviors .



① : For each site, select the pairs of AA & WA which were matched in the confounding and behavioral variables. The matching was performed at the subject level, rather than the group level.

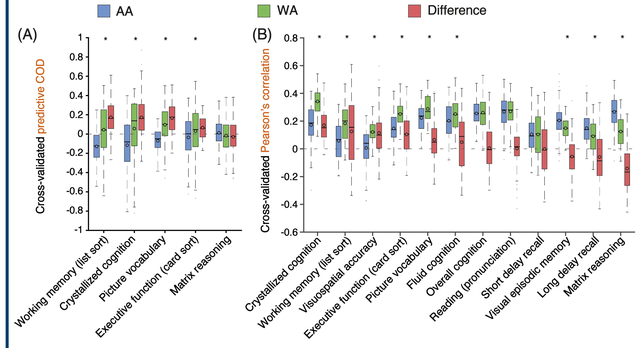
② : Merge 19 sites into 10 sets so that # matched AA were as balanced as possible across sets.

③ : Select 3 sets as test folds (red bounding box), the remaining 7 sets as training folds, yielding 120 possible data splits.

- **Kernel ridge regression** (Kong 2019; Li 2019; He 2020):
  - The behavior of a test subject is more similar to the behavior of a training subject if their brain organizations are more similar.
  - Inter-subject similarity (i.e. kernel): correlation of subjects' RSFC matrices.
  - 120 variations of cross-validated training-test data split.
- **Accuracy metrics:**
  - Predictive COD (AA as example, similar for WA): $pCOD_{AA} = 1 - \frac{SSE_{AA}}{SST_{AA\&WA}}$, where
    $SSE_{AA} = \sum(\text{AA test predicted score} - \text{AA test true score})^2$
    $SST_{AA\&WA} = \sum(\text{matched AA\&WA training true score} - E[\text{matched AA\&WA training true score}])^2$
    Assumption: total data variance is not group specific
  - Pearson's correlation
- **Brain-behavioral association (BBA; Haufe 2014)**
  - Model-learned BBA: covariance[RSFC, predicted behavioral scores] across training subjects
  - True BBA in each ethnic/racial group (either AA or WA): covariance[RSFC, true behavioral scores] across test subjects in that group.

Out-of-sample prediction accuracies biased towards WA

- No significant difference in confounding and behavioral variables was found between matched AA and WA (FDR q < 0.05).
- Confounding variables (age, gender, FD, DVARS, ICV, parental education) were regressed from behavioral scores before building predictive models. Data leakage from training to test sets were prevented.
- Most predictable behavioral measures[1] showed significantly higher prediction accuracy in matched WA than AA[2].



[1] "Predictable behavior" satisfied 3 conditions: (1) Pearson's correlation accuracy of all test subjects > 0.15 (2) survived the permutation test by shuffling the predicted scores across all test subjects (FDR q < 0.05); (3) prediction accuracy is positive in either AA or WA.

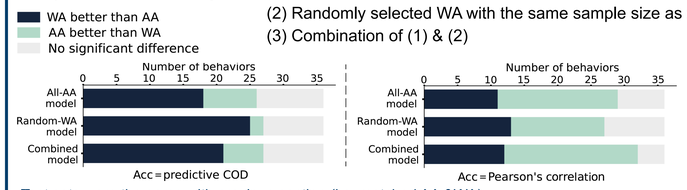
[2] AA vs WA accuracy difference tested by randomly shuffling AA/WA labels 1000 times, FDR q < 0.05

* AA vs WA accuracy difference were significant.

- When the same confounding variables were regressed from RSFC, similar conclusion can be drawn.

Effects of training population

Train predictive models only on: (1) All AA;
(2) Randomly selected WA with the same sample size as (1);
(3) Combination of (1) & (2)

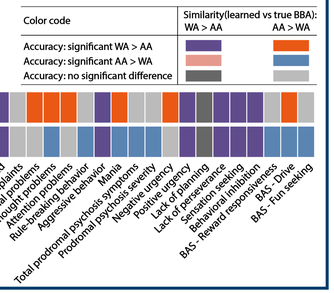


Test sets were the same with previous section (i.e. matched AA &WA).

- Across all 36 behavioral measures, accuracies of AA became less disadvantaged when models were trained solely on AA, compared to the other two types of models.
- However, training solely on AA did not eliminate the accuracy differences between AA and WA.

Models learned brain-behavioral association patterns more similar to true patterns of WA than that of AA

Behavioral measures showing more similar model-learned BBA patterns and actual BBA patterns in one ethnic/racial group than the other tended to have greater accuracy in the same group (#purple+#blue > #pink+#orange).



Discussion

1. When predictive models were trained on entire sample (dominated by WA), significantly lower out-of-sample accuracies in AA than WA were observed for most predictable behavioral measures.
   The results replicated our previous study on young, healthy adults using the Human Connectome Project dataset (Li 2020).
2. The observed bias in model performance was related to the similarity between model-learned brain-behavioral association (BBA) patterns and the actual BBA patterns in different ethnic/racial groups.
3. Training solely on AA could benefit the prediction in AA, compared to specific training on only WA.
   However, some significant AA-WA differences in the test accuracies were still observed.
4. The above observation raised further questions on other possible factors contributing to the difference in prediction difference which need future investigations, e.g. choice of brain templates during preprocessing, validity of psychometric data in minority ethnic/racial groups (Gould 1996).
5. Call for data collection from non-European descended populations: most of current large-scale neuroimaging + behavior datasets are dominated by European/white subjects. E.g. UK Biobank, the currently largest dataset, only contains 1% Asian British (N~=300, the largest non-European/white group) with both fMRI and behavioral data.
6. Given complicatedly entangled nature and nurture in present data, this study should not be interpreted as identifying neurobiological and neuropsychological differences across groups which would potentially lead to more discriminations.

References

[1] Chen et al., (2020). Shared and unique brain network features predict cognition, personality and mental health in childhood. bioRxiv, p. 2020.06.24.168724
[2] Chouldechova A, Roth A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.
[3] Finn ES et al., (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neuroscience. 18(11):1664.
[4] Fischl B et al., (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 33:341-55.
[5] Gould, SJ. et al. (1996). The mismeasure of man. WW Norton & company
[6] Haufe, S. et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87:96-110.
[7] He, T. et al., Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics, NeuroImage, https://doi.org/10.1016/j.neuroimage.2019.116276
[8] Li J et al., (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. NeuroImage. 196:126-41.
[9] Li, J. et al. (2020). Unfairness in RSFC-based behavioral prediction across African American & White American Samples [Abstract]. 2020 Organization of Human Brain Mapping Annual Meeting
[10] Kong R et al. (2019). Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. Cerebral Cortex. 29(6):2533.
[11] Martin AR et al., (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics. 51(4):584-91.
[12] Obermeyer Z, et al., (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science. 366(6464):447-53.
[13] Schaefer A et al., (2017). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex. 28(9):3095-3114.

Acknowledgments

S.G. is supported by the Heisenberg Programme of the Deutsche Forschungsgemeinschaft (GE 2835/2–1), J.L. S.B.E. and K.R.P. are supported by the Deutsche Forschungsgemeinschaft (EI 816/4–1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain", the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements No. 945539 (HBP SGA3) and 826421 (VirtualBrainCloud). B.T.T.Y., J.C., A.T. and L. Q.R.O. are supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017) and the Singapore Ministry of Defense (Project CURATE). Our research also utilized resources provided by the Center for Functional Neuroimaging Technologies, NIH P41EB015896 and instruments supported by NIH 1S10RR023401, NIH 1S10RR019307, and NIH 1S10RR023043 from the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital. Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).