Unfairness in RSFC-based behavioral prediction across African American & White American Samples

Heinrich Heine Universität Düsseldorf

NUS National University of Singapore

#1275

Jingwei Li[1,5], Danilo Bzdok[2,3], Avram Holmes[4], B.T. Thomas Yeo[5], Sarah Genon[1,6]

[1]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany, jin.li@fz-juelich.de
[2]McGill University, Canada [3]Mila - Quebec Artificial Intelligence Institute, Canada [4]Yale University, USA [5]ECE, CSC, CIRC, N.1 & MNP, NUS, Singapore [6]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Germany

JÜLICH Forschungszentrum

Brain & Behaviour

Introduction

While machine learning will likely play a major role in precision medicine, there are growing concerns that machine learning algorithms might exhibit unfairness against under-represented and other sub-populations (Chouldechova 2018; Martin 2019; Obermeyer 2019).

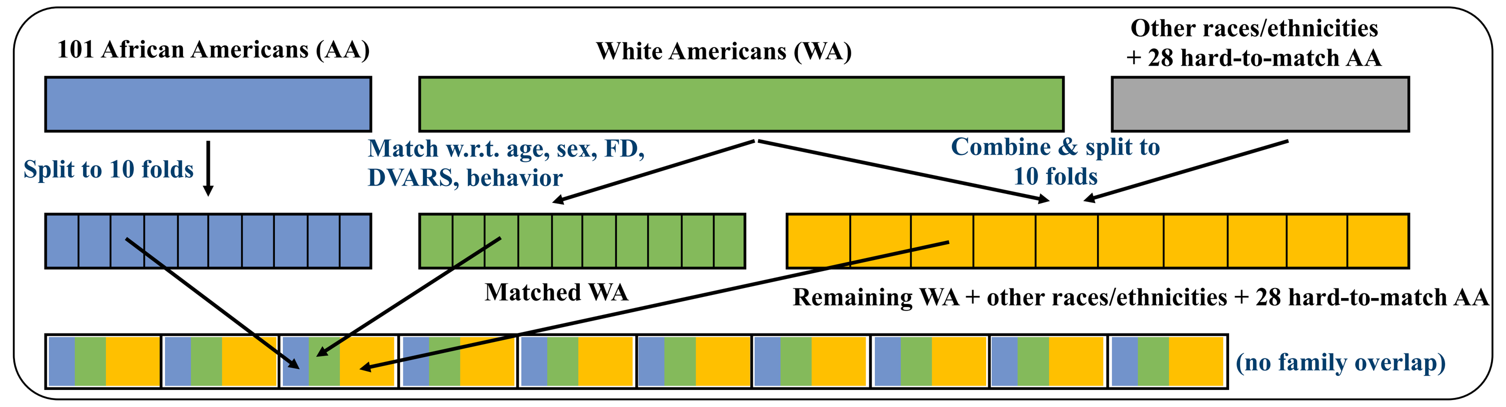
Given significant interests and efforts in predicting behavioral phenotypes with resting-state functional connectivity (RSFC; Finn 2015), here, we examined potential differences in RSFC-based behavioral prediction performance between African American (AA) and matched White American (WA) samples.

Different conclusions were drawn using various accuracy metrics. Using predictive COD, behavioral prediction model trained on our entire population exhibited significantly worse performance in AA compared with matched WA for most behaviors examined. However, some behaviors showed higher Pearson's correlation accuracy in AA than WA. The inconsistency could be partially due to the higher behavioral variance and higher prediction shift in AA than matched WA. We encourage more data for minorities to be collected, to better understand the reasons causing different model performances among the subpopulations.

Methods

- **Dataset:** Human Connectome Project (HCP; Van Essen 2013; Smith 2013) S1200 release (N = 948, incl. 129 African Americans, 721 White Americans, 58 behaviors)
- **RSFC preprocessing:**
  - ICA-FIX (Salimi-Khorshidi 2014) + global signal regression (Li 2019)
  - RSFC across 400 cortical (Schaefer 2018) & 19 subcortical (Fischl 2002) ROIs.
- **101 pairs** of AA & WA were obtained after **matching** for age, sex, FD, DVARS & behavior.
  - Education, household income, intracranial volume cannot be matched without excluding large number of AA subjects.
  - 7/58 behaviors cannot find matched WA for enough (i.e. 40) random splits of AA.

Data split & matching, performed for each individual behavior



- **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.
  - Nested 10-fold cross-validation, randomly repeated 40 times.
- **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
  - Normalized MSE (AA as example, similar for WA):
    $normMSE_{AA} = MSE_{AA}/var[\text{AA training true score}]$
    $normMSE_{WA} = MSE_{WA}/var[\text{matched WA training true score}]$

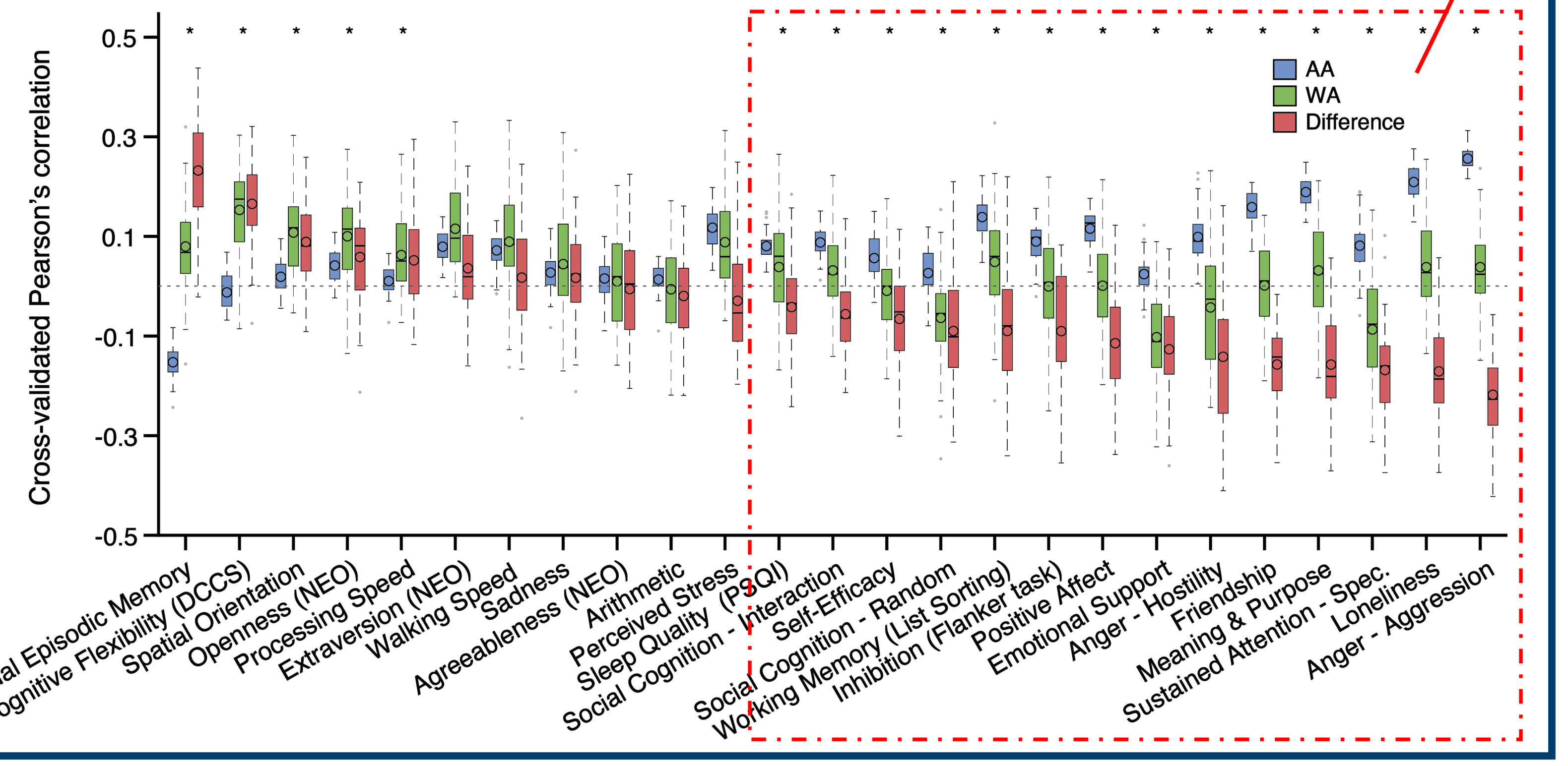
AA-WA differences vary using different accuracy metrics

- All analyses focused on the 101 matched AA & WA groups. No significant difference was found between the two groups for the 5 matching variables (FDR q < 0.05).
- The same confounding variables were regressed from either behaviors or RSFC: age, sex, FD, DVARS, education, household income, intracranial volume.

Accuracy metric:	Predictive COD		Pearson's correlation	
Regress covariates from:	Behaviors	RSFC	Behaviors	RSFC
# behaviors predictable[1]	29	23	32	25
# behaviors with significant AA vs WA accuracy difference[2]	26 (WA>AA)	22 (WA>AA)	28 10 (WA>AA) 18 (AA>WA)	19 5 (WA>AA) 14 (AA>WA)

[1] "Predictable behavior": survived the permutation test by shuffling the predicted scores across subjects (FDR q< 0.05), and the predictive COD (or Pearson's correlation) value is positive in either AA or WA.
[2] Permutation test by shuffling AA/WA labels, FDR q < 0.05

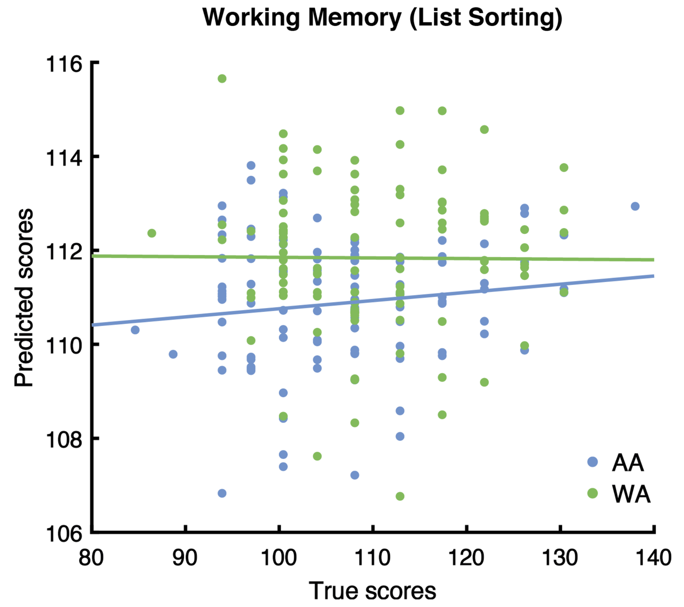
AA-WA difference in Pearson's correlation for individual predictable behaviors
(* indicates significant AA-WA difference.)



Higher behavioral variance & prediction shift in AA than matched WA

Possible reasons of inconsistency between predictive COD & Pearson's correlation:

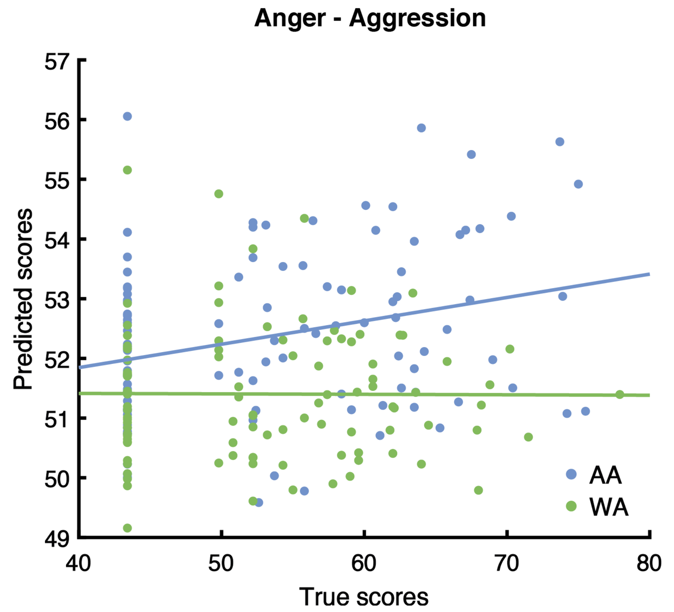
1. Overall shift of predicted scores, i.e. $(E[\textbf{predicted score}] - E[\textbf{true score}])^2$ cannot be captured by correlation, e.g.:
2. Variance of true behavioral scores: AA > WA, e.g.:



Pearson's correlation:
- AA: 1.3
- WA: -0.0080

Predictive COD:
- AA: -0.22
- WA: 0.053

Overall shift: • AA: 18 • WA: 3.7



Pearson's correlation:
- AA: 0.28
- WA: -0.0058

Predictive COD:
- AA: -0.12
- WA: 0.11

Overall shift: • AA: 2.6 • WA: 2.8

Variance of true scores: • AA: 99 • WA: 72

Generally, we observed higher overall prediction shift and higher variance of true behavioral scores in AA than matched WA.

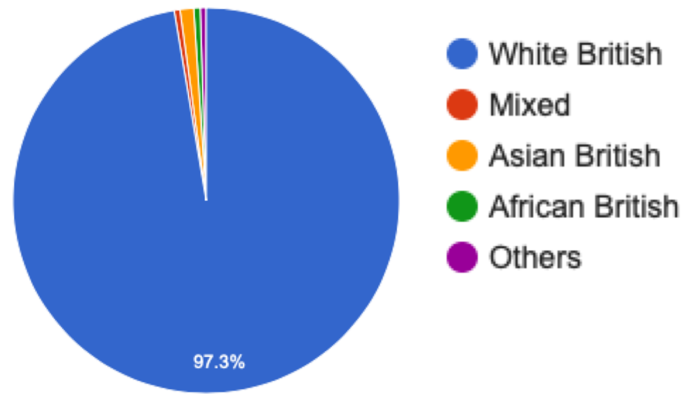
Fewer behaviors showed significant accuracy difference using normalized MSE as the metric (i.e. consider AA-WA difference in behavioral score variance):

Regress covariates from:	Behaviors	RSFC
# behaviors predictable (using predictive COD)	29	23
# behaviors with significant AA vs WA accuracy difference	10 8 (WA>AA); 2 (AA>WA)	6 3 (WA>AA); 3 (AA>WA)

Discussion

1. Perfect matching for some demographic / morphologic / behavioral variables was NOT possible in current data. The current strategy was to regress them from behaviors or functional connectivity.
2. Models trained on full population predicted AA & WA differently, even after regressing confounding variables such as education, income and intracranial volume. One possibility is that there are other confounding variables beyond the ones we examined here. Another reason could be that the influence of these variables is not linear.
3. In the maximally matched samples, AA showed higher behavioral variance than WA. The difference in behavioral variance further affected the accuracy metrics.
4. To better study the performance of behavioral prediction models in different subpopulations, better matching between the subpopulations is needed. Hence more data for the minorities need to be collected.
5. We will explore this question using other datasets like UK-Biobank and NKI, but the data for minorities may be still not enough. For example in UK-Biobank, the largest minor ethnicity, Asian British, occupies only 1% of total sample size (N ~= 300 with both RSFC and cognitive behavioral data before quality control).

Ethnicities in UK-Biobank



References

[1] Chouldechova A, Roth A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.
[2] Finn ES et al., (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neuroscience. 18(11):1664.
[3] Fischl B et al., (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 33:341-55.
[4] 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
[5] Li J et al., (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. NeuroImage. 196:126-41.
[6] Kong R et al. (2019). Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. Cerebral Cortex. 29(6):2533.
[7] Martin AR et al., (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics. 51(4):584-91.
[8] Obermeyer Z, et al., (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science. 366(6464):447-53.
[9] Schaefer A et al., (2017). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex. 28(9):3095-3114.
[10] Smith, S.M. et al. (2013). Resting-state fMRI in the human connectome project, Neuroimage, 80:144-168.
[11] Van Essen, D.C. et al., (2013) Wu-Minn HCP Consortium. The WU-Minn human connectome project: an overview. Neuroimage. 15;80:62-79.

Acknowledgments

This study was supported by Singapore MOE Tier 2 (MOE2014-T2-2-016), NUS Strategic Research (DPRT/944/09/14), NUS SOM Aspiration Fund (R185000271720), Singapore NMRC (CBRG/0088/2015), NUS YIA and the Singapore National Research Foundation (NRF) Fellowship (Class of 2017). 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). It was also supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/1-1, EI 816/4-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme 'Supercomputing and Modeling for the Human Brain', and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1) and No. 785907 (HBP SGA2).