Characterizing Data-Driven Phenotypes of Schizophrenia and ADHD Using the Consortium for Neuropsychiatric Phenomics (CNP) Dataset

Scott L. Fleming
scottyf@stanford.edu

Motivation and Overview

- Psychiatrists have traditionally relied on the *Diagnostic Statistical Manual of Mental Disorders* (DSM-5) to make diagnoses.
- The DSM-5 has been criticized because boundaries between disorders are not as strict as it suggests [1].
- An ongoing challenge in the field is to identify disease biomarkers that correlate with underlying brain dysfunction [2].
- Schizophrenia and ADHD in particular share core features (e.g. attention dysfunction) but are distinct categories in the DSM-5 [3].
- This research uses 1) Hierarchical clustering, 2) PCA, and 3) classification via Multinomial Regression on a Neuropsychiatric Phenomics dataset [4] to answer the following questions – *how distinct are Schizophrenia and ADHD? Is it possible to distinguish them based on objective neurological features alone?*

Data

- A dataset shared on OpenfMRI, from UCLA's Consortium for Neuropsychiatric Phenomics, with 1,998 measurements on 272 participants including demographics, symptoms/traits, and 556 features from neurological/neurocognitive tasks [4].
- Ground truth labels are professional diagnosis of schizophrenia, ADHD, bipolar disorder, or "healthy".

Domain	Measure
Demographics & General Health	• Study-specific demographic form • General Health Questionnaire • Smoking Status
Symptoms	• Young Mania Rating Scale-C (YMRS)* • Hamilton Psychiatric Rating Scale for Depression (HAMD-28)* • Scale for the Assessment of Negative Symptoms (SANS), Scale for the Assessment of Positive Symptoms (SAPS)* • Brief Psychiatric Rating Scale (BPRS)* • Hopkins Symptom Checklist (HSCL) • Adult Self-Report Scale Screener (ASRS)
Traits	• Barratt Impulsiveness Scale (BIS-11) • Dickman Functional and Dysfunctional Impulsivity Scale • Multidimensional Personality Questionnaire (MPQ)—Control subscale • Impulsiveness, Venturesomeness and Empathy Scale (IVE) • Scale for Traits that Increase Risk for Bipolar II Disorder • Golden & Meehl's Seven MMPI Items Selected by Taxonomic Method • Hypomanic Personality Scale (HPS) • Chapman Scales (Perceptual Aberrations, Social Anhedonia, Physical Anhedonia) • Temperament & Character Inventory (TCI) • Munich Chronotype Questionnaire (MCTQ)
Neurocognitive Tasks	• Task-switching Task • Spatial Capacity Task (SCAP) • Verbal Capacity Task (VCAP) • Delay Discounting Task (DDT) • Attention Network Task (ANT) • Continuous Performance Go/NoGo (CPT) • Stroop Color Word Task (SCWT) • Stop Signal Task (SST) • Scene Recognition Task
Neuropsychological Assessment	• California Verbal Learning Test (CVLT-II) • WMS-IV Symbol Span • WMS-IV Visual Reproduction • WAIS-IV Letter Number Sequencing • WMS-IV Digit Span • WAIS-IV Vocabulary • WAIS-IV Matrix Reasoning Color Trails Test

* = Assessment given to patients with diagnosed disorder only (not given to healthy controls)

Features and Data Splitting

- After removing rows (patients) and columns (features) with high missingness, remaining missing values were imputed with KNN-imputation, leaving 1270 features total, with 330 "objective" neurocognitive/neuropsychiatric features.
- Used PCA to reduce features space down to 77 features for disorders-only analysis, 154 features for analysis including healthy controls
- Split the data into train, validation, and test set

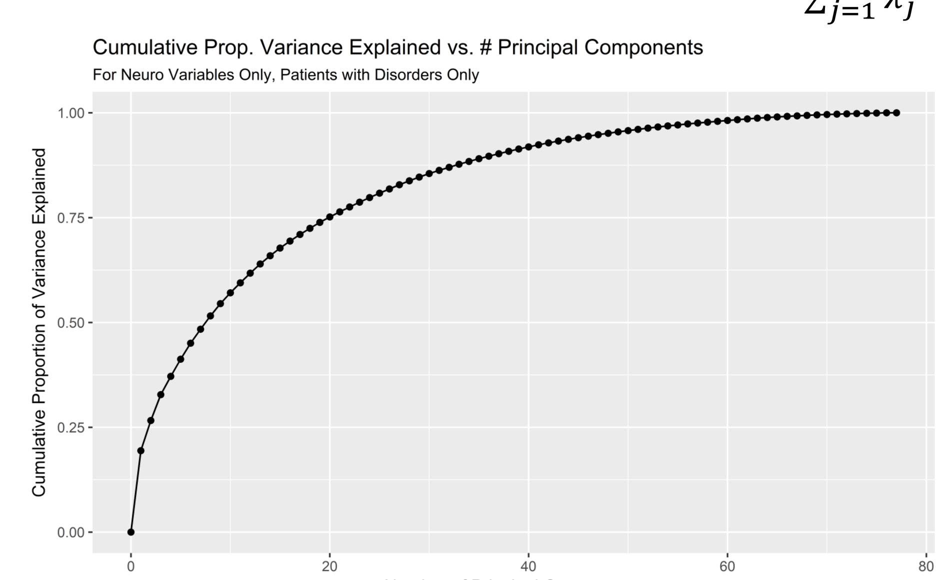
TABLE I: *Breakdown of the CNP Cohort*

Variable Name	Original Dataset	Train Set	Validation Set	Test Set
Control	130	56	24	19
ADHD	43	19	7	6
Bipolar	49	23	9	8
Schizophrenia	50	21	8	7
Total	**272**	**119**	**48**	**40**

Models (PCA)

- First k principal components can be found using the top k eigenvectors of Σ. So if we let λ_i be the i^{th} eigenvalue of Σ, then…

$$\text{Proportion of Variance Explained by PC } i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$$



- Additionally, we used loadings matrix to get a percent contribution of each variable to each PC:

$$\text{Loadings Matrix}(X_i, PC_j) = \alpha_{i,j}$$

$$\%\text{ Contrib. of } X_i \text{ to } PC_j = \frac{|\alpha_{i,j}|}{\sum_{i=1}^{n} |\alpha_{i,j}|}$$

- See plots above right for visualization of top 10 variables (by % contribution) for first two PCs

Models (Hierarchical Clustering)

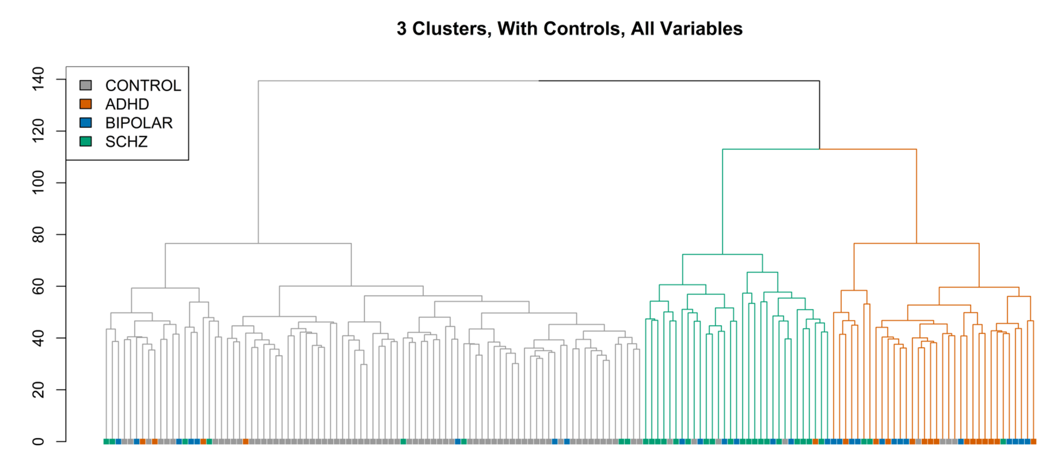
- Used hierarchical agglomerative clustering with Ward's linkage [5] and squared Euclidean distance.
- Used the Gap Statistic to choose the "optimal" number of clusters [6] (first maximum criterion).
- If k = 1, 2, …, K is the cluster label, D_r is the sum of all pairwise distances in cluster r, then

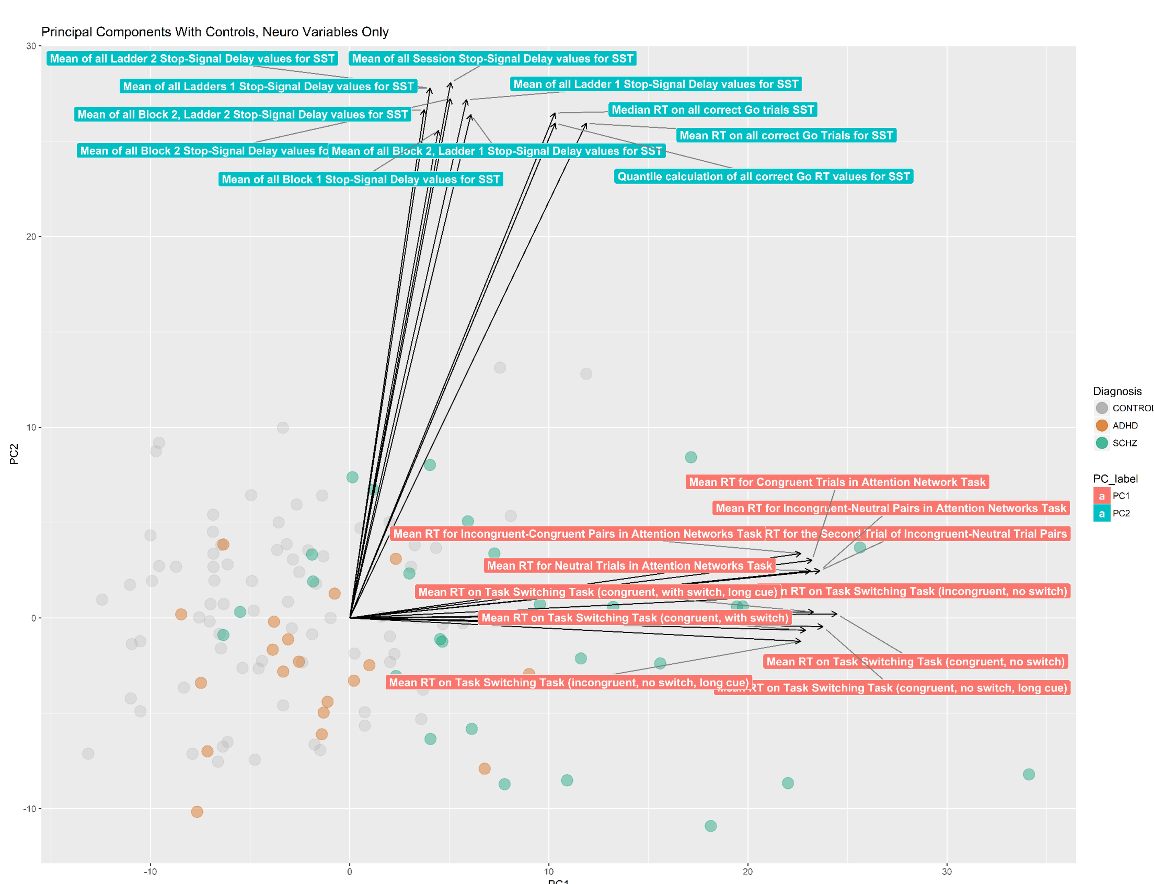
$$W_k = \sum_{r=1}^{k} \frac{1}{2|C_r|} D_r$$

$$\text{Gap}_n(k) = (1/B) \sum_b \log(W_{kb}^*) - \log(W_k)$$

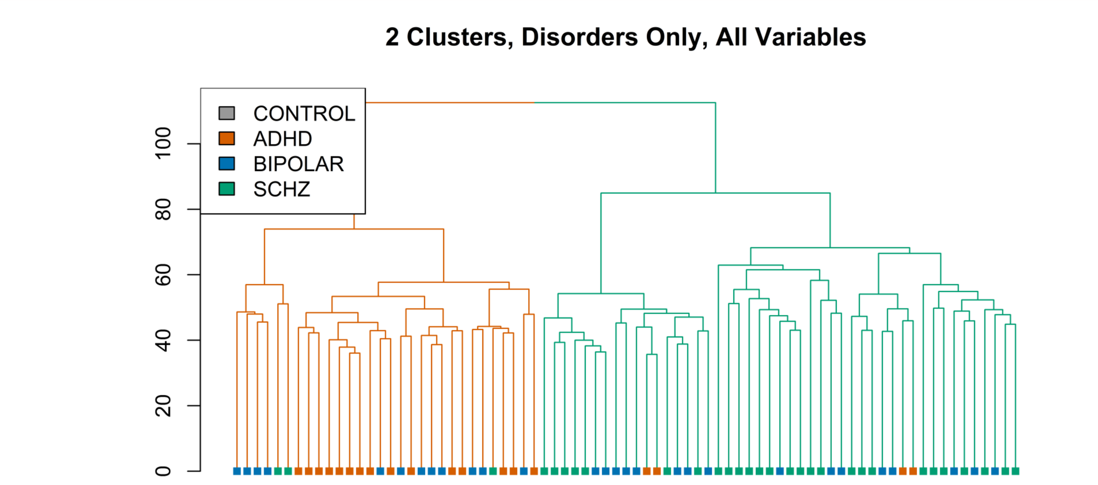
Where W_{kb}^* is simulated from a uniform null distribution.

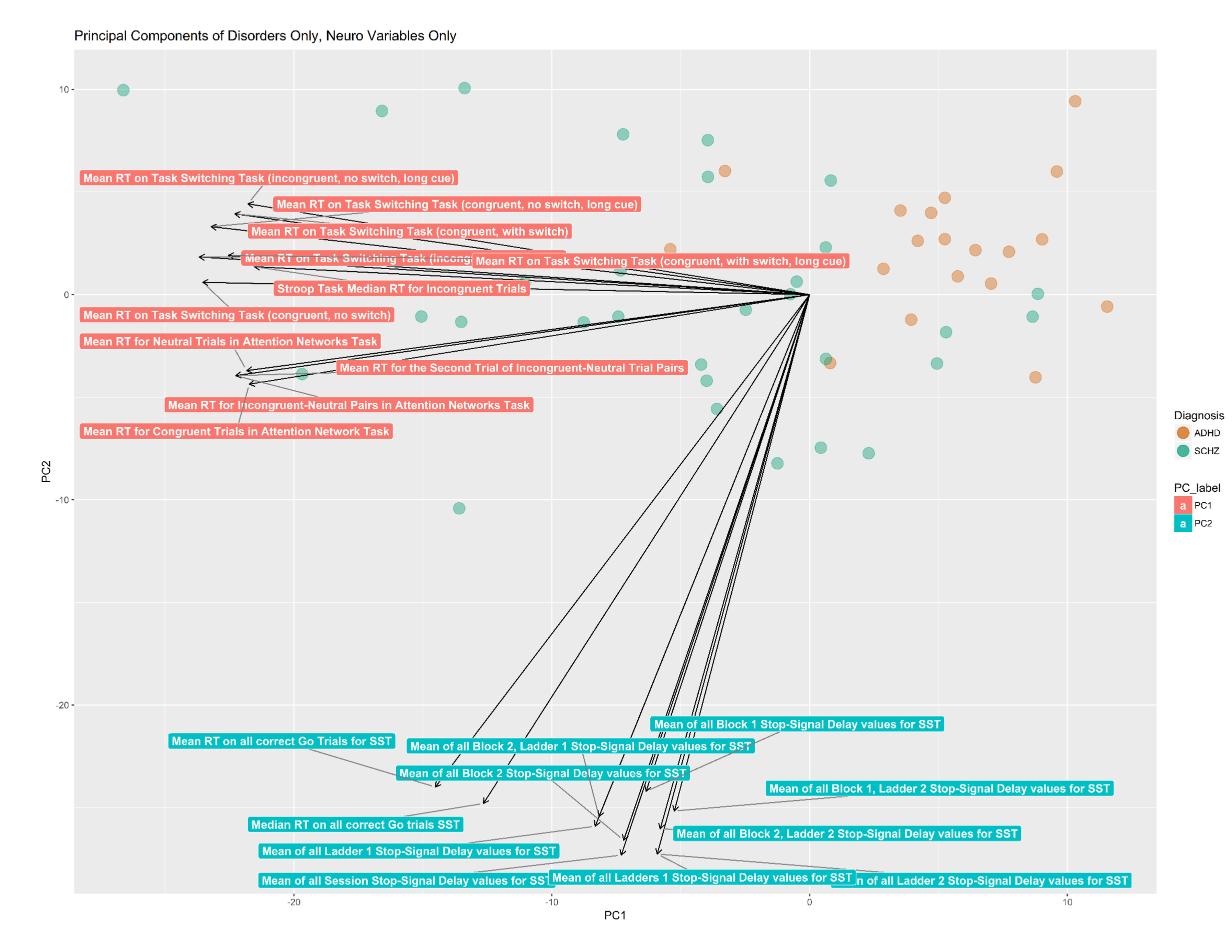
Results (With Controls)





Results (Disorders Only)





Results (Multinomial Regression, Using Neuro Variables Only)

The Model

- Modeled the response function as a Multinomial (i.e. outcome variable G has K levels in {"CONTROL", "ADHD", "SCHZ"}) [7]:

$$\Pr(G = k | X = x) = \frac{e^{\beta_{0k} + \beta_k^T x}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell} + \beta_\ell^T x}}.$$

- Used regularized regression/lasso penalty to encourage sparsity.

With Controls

- Overall validation accuracy of 63.16%
- ADHD sensitivity = 0, specifity = 0.90
- SCHZ sensitivity = 0.5, specificity = 0.90

		True Labels		
		CONTROL	ADHD	SCHZ
Predicted Labels	CONTROL	21	7	1
	ADHD	1	0	2
	SCHZ	3	0	3

Disorders Only

- Overall validation accuracy of 84.62%
- ADHD sensitivity = 0.85, specifity = 0.83
- SCHZ sensitivity = 0.83, specificity = 0.85

		True Labels	
		ADHD	SCHZ
Predicted Labels	ADHD	6	1
	SCHZ	1	5

Discussion

- Good performance of our classifier on the validation set using just measurements from neurocognitive tasks and neuropsychological assessments suggests underlying neurological basis for the DSM-5 categories of Schizophrenia and ADHD
- "Most important" features from PCA seem to be related to attention, which confirms previous discussions of attention being a core construct in both ADHD and Schizophrenia [3].
- Poor performance distinguishing controls from disorders merits further exploration

Future Analysis

- Explore fMRI imaging features as well
- Try different classifiers to see if model with controls improves in performance

References and Acknowledgements

1. Casey, B. J., et al. "DSM-5 and RDoC: progress in psychiatry research?." *Nature Reviews Neuroscience* 14.11 (2013): 810-814.
2. Drysdale, Andrew T., et al. "Resting-state connectivity biomarkers define neurophysiological subtypes of depression." *Nature medicine* 23.1 (2017): 28-38.
3. Pallanti, Stefano, and Luana Salerno. "Raising attention to attention deficit hyperactivity disorder in schizophrenia." *World journal of psychiatry* 5.1 (2015): 47.
4. Poldrack, R. A., et al. "A phenome-wide examination of neural and cognitive function." *Scientific data* 3 (2016): 160110.
5. Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function." *Journal of the American statistical association*58.301 (1963): 236-244.
6. Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001): 411-423.
7. Hastie, Trevor, and Junyang Qian. "Glmnet vignette." (2014).

***Special thanks to Dr. Russell Poldrack for guiding me, to Dr. Leanne Williams and the NIH for funding me, and to Dr. Matthew Sacchet for insight from a Cognitive Scientist's Perspective**