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 OpenMNI, from UCLA’s Consortium for Neuropsychiatric Phenomics, with 1,998 measurements on 272 participants including demographics, symptoms/trait, and 556 features from neurological/neurocognitive tasks [4].
- Ground truth labels are professional diagnosis of schizophrenia, ADHD, bipolar disorder, or “healthy”.

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 disorder-only analysis, 154 features for analysis including healthy controls
- Split the data into train, validation, and test set

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...

\[
\text{Proportion of Variance Explained by PC } i = \frac{\lambda_i}{\sum_{j=1}^{k=77} \lambda_j}
\]
- Additionally, we used loadings matrix to get a percent contribution of each variable to each PC:

\[
\text{Loadings Matrix}(X_{p,c}) = \frac{d_{i,j}}{\sum_{i=1}^{n} \sqrt{d_{i,j}^2}}
\]

\[
\%\text{ Contrib. of } X_i \text{ to } PC_c = \frac{|d_{i,j}|}{\sum_{j=1}^{77} |d_{i,j}|}
\]
- See plots above right for visualization of top 10 variables (by % contribution) for first two PCs

Models (Hierarchical Clustering)

\[
\text{Distance between clusters } D_{ij} = \frac{\sum_{k=1}^{K} \text{min}(d_{ik}, d_{kj})}{\min(K-1, K)}
\]
- Useplots below to visualize the top 100 clusters (by % contribution) for first two PCs

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 distinguishes controls from disorders merits further exploration

References and Acknowledgements
1. "A dataset shared on OpenMNI, from UCLA’s Consortium for Neuropsychiatric Phenomics, with 1,998 measurements on 272 participants including demographics, symptoms/trait, and 556 features from neurological/neurocognitive tasks [4].
2. "Ground truth labels are professional diagnosis of schizophrenia, ADHD, bipolar disorder, or “healthy”.

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