Submitter Name: Jinbo Bi
Submitted Email: jinbo.bi@uconn.edu

Inferring Phenotypes from Substance Use via Collaborative Matrix Completion

Jin Lu[1], Jiangwen Sun[1], Xinyu Wang[1], Henry R Kranzler[2], Joel Gelernter[3], Jinbo Bi[1]

[1]Department of Computer Science and Engineering, University of Connecticut; [2]Center for Studies of Addiction, University of Pennsylvania Perelman School of Medicine; [3]Department of Psychiatry, Yale University

Background: Researchers often aggregate subjects from multiple studies in genetic analysis, but these subjects can have missing diagnostic criteria for certain substances that were not originally a focus of study. Recent advances in addiction neurobiology have shown that the abuse of multiple substances has similar genetic determinants.

Hypothesis: The shared biological base of concurrent disorders can help infer missing diagnostic criteria of a disorder from known criteria of another comorbid disorder.

Method: We performed genome-wide association analyses (GWA) with the cocaine use disorder (CUD) and opioid use disorder (OUD). We then developed a new approach to infer unreported diagnostic criteria of CUD from the correlated OUD criteria and vice versa. The top genetic markers identified by the GWA were used in the inference. It was tested on 3,441 subjects who had both CUD and OUD to evaluate the inference accuracy and identify the most important gene-symptom interactions useful for the inference.

Results: The GWA identified 383 genetic variants with p-value < 10e-5. Based on these genetic variants, and 11 CUD and 11 OUD criteria defined by DSM-V, our method achieved an accuracy (RMSE~0.23), much better than the state of the art (RMSE~0.3). It identified many gene-symptom interactions. For example, the SNP rs1481605 at base pair 13,519,829 on chromosome 8, located at the downstream of gene *C8orf48*, received the highest weights for its interactions with all 22 criteria in the inference model.

Conclusion: The proposed phenotype inference method is a promising tool to impute unreported or unobserved criteria for disease diagnosis.