Collaborative Phenotype Inference from Comorbid Substance Use Disorders and Genotypes

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

Data in large-scale genetic studies of complex human diseases, such as substance use disorders, are often incomplete. Despite great progress in genotype imputation, e.g., the IMPUTE2 method, considerably less progress has been made in inferring phenotypes. We designed a novel approach to integrate individuals' comorbid conditions with their genotype data to infer missing (unreported) diagnostic criteria of a disorder. The premise of our approach derives from correlations among symptoms and the shared biological bases of concurrent disorders such as co-dependence on cocaine and opioids. We describe a matrix completion method to construct a bi-linear model based on the interactions of genotypes and known symptoms of related disorders to infer unknown values of another set of symptoms or phenotypes. An efficient stochastic and parallel algorithm based on the linearized alternating direction method of multipliers was developed to solve the proposed optimization problem in a large scale. The approach was evaluated and compared against several other advanced data matrix completion methods in a case study. A total of 7,189 subjects were aggregated from family and case-control based genetic studies of cocaine use disorder (CUD) and opioid use disorder (OUD). The CUD (or OUD) criterion count was derived by counting the number of criteria endorsed by an individual out of the eleven DSM-5 criteria and was used in a GWAS to identify genetic variants.

Our results show that the proposed method outperformed other methods in terms of imputation accuracy measured by the root mean squared error even when the run time of our method was significantly reduced by nearly 95% of non-stochastic methods. Our method could readily handle big data due to greater computational efficiency. The analysis using the proposed method identified several gene-clinical interactions that are substantially useful for inferring SUD criteria. In particular, the interactions between markers at 9:136198589, 8:13519129, 8:13517808, 8:13517469, 8:13520905, 8:13530761, 8:13571774, and all the criteria defined in DSM-5 for CUD and OUD had the largest effects in the inference model for the missing criteria. These markers were also among the top 10 markers identified in a genome-wide association study of CUD and OUD. These results indicate that using both additive effects of genetic markers and observed clinical symptoms can more effectively estimate or infer unreported SUD criteria, and the proposed method is promising to perform this task.