Machine learning to identify highly heritable components of substance use disorders

Jinbo Bi[1], Jiangwen Sun[1], Henry R Kranzler[2]

Substance use disorder (SUD) is clinically heterogeneous and a SUD diagnosis is based on multiple clinical criteria. Identifying highly heritable components or subtypes of SUD could maximize the likelihood of finding genetic associations. Existing methods for refinement of disease phenotypes perform unsupervised cluster analysis on clinical criteria and hence do not assess heritability. Existing heritable component analytics either cannot utilize general pedigrees or have to estimate a large covariance matrix of clinical features from limited samples, which leads to inaccurate estimates and is often computationally prohibitive. These methods are also difficult to correct the fixed effects from covariates such as age or sex to identify truly heritable components. The proposed approach searches for a combination of clinical features and directly maximizes the heritability of this combined trait. A quadratic optimization problem is derived where the objective function is formulated by decomposing the maximum likelihood method for heritability estimation. This new approach is computationally efficient and generates linearly-combined traits of higher heritability than those by other methods, with correction for fixed effects. Using 6,810 subjects with 13 cocaine use and related behavioral variables, this method yielded a subtype of cocaine use disorder (CUD) with early first-use of cocaine, early onset of dependence, heavy cocaine use, and an estimated heritability of 0.7. A genome-wide association study, that compared the utility of the derived subtype with the commonly used CUD phenotype, identified more statistically significant associations with the derived subtype with replication.

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