Application number/Title: 40404- Topic Modeling for Phenomewide Studies in Biobanks

Applicant PI: Dr Roy Perlis

Application Institution: Massachusetts General Hospital, Boston, Massachusetts, USA

Keywords provided by the Applicant PI to describe the research project:

cluster analysis, coded clinical data, genetic association, latent Dirichilet allocation, machine learning, reverse genetics

Application Lay Summary:

Biobanks and national registries represent a powerful tool for genomic discovery, but rely on diagnostic codes that may be unreliable and fail to capture the relationship between related diagnoses. We are proposing to apply a new method that identifies and groups related disorders for study. In preliminary studies in another biobank, we have demonstrated that this method makes better use of diagnostic codes than studies that simply examine every single code individually. We will use this new method to identify genetic variations that contribute, not to single diseases, but to groups of diseases. From a public health perspective, this study will help researchers understand how seemingly different diseases are related, at a clinical as well as genetic level, which may help us understand the causes of these diseases. We anticipate completing this project within 18-24 months.