The identification of comorbidity risk via disease-disease network: an application to pre-eclamptic women in the UK Biobank

Vivek Sriram1,2, Seung Mi Lee2,3, Yonghyun Nam2, Dokyoon Kim2
1Genomics and Computational Biology Graduate Group, University of Pennsylvania; 2Department of Epidemiology, Biostatistics and Informatics, University of Pennsylvania; 3Department of Obstetrics and Gynecology, Seoul National University Hospital

Background

- Pre-eclampsia is a hypertensive disease that occurs during pregnancy – can lead to exacerbated health outcomes and increased comorbid risk
- A disease-disease network (DDN), a graph where nodes represent phenotypes and edges represent SNPs shared between phenotypes, can help visualize the genetic relationships across diseases
- In this study, we apply graph-based machine learning to identify novel comorbidity correlations and rank phenotypes according to their genetic similarity to pre-eclampsia

Study Dataset

- We applied our method to SAIGE-analyzed UK Biobank PheWAS summary data:
  - 1400 phenotypes
  - 28 million variants
  - 400,000 White British individuals

Network Generation

A network depicting the associations between phenotypes in our UKBB dataset – brown nodes represent source pre-eclamptic phenotypes, blue nodes represent associated diseases according to PheWAS summary data, orange nodes represent associated diseases according to UKBB case occurrences, and purple nodes represent associated diseases according to both PheWAS and UKBB.

- In order to consider SNPs to be significantly associated with phenotypes, we established a minor allele frequency threshold of 0.05, a case count of 200, and a p-value of 1e-4
- SNPs were then LD-pruned with an R-squared threshold of 0.2
- All nodes related to injuries or symptoms were dropped from our network
- Our final network included 865 nodes and 62,111 edges

Graph-based Machine Learning Inference

- Graph-based Semi-Supervised Learning (GBSSL) is a machine learning approach for signal propagation according to the topology of a network
- GBSSL allows us to assign labels of “+1” to an initial set of phenotypes, and then determine relative labels on a range from 0 to 1 for all other phenotypes

Results

- Chi-square tests yield 65/871 (7.46%) associations with preeclampsia
- GBSSL with elbow thresholding suggests 91 diseases to be associated with pre-eclampsia – 10 of these are positives according to Chi-square
- GBSSL confirms previously identified diseases (i.e. preeclampsia, abruptio placenta, and hemorrhage during pregnancy) and suggests novel ones (i.e. subarachnoid hemorrhage and abnormal findings of biliary tract) as comorbid with preeclampsia

Summary

- GBSSL applied to SNP-based DDNs offers an effective way to identify potential disease comorbidities and uncover the genetic architecture of disease connections
- Future works include developing alternative DDNs that capture clearer genetic associations between diseases for improved comorbidity inference