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

Vivek Sriram[1,2], Seung Mi Lee[2, 3], Yonghyun Nam[2], Dokoyoon Kim[2]

[1]Genomics and Computational Biology Graduate Group, University of Pennsylvania; [2]Department of Epidemiology, Biostatistics and Informatics, University of Pennsylvania; [3]Department of Obstetrics and Gynecology, Seoul National University Hospital

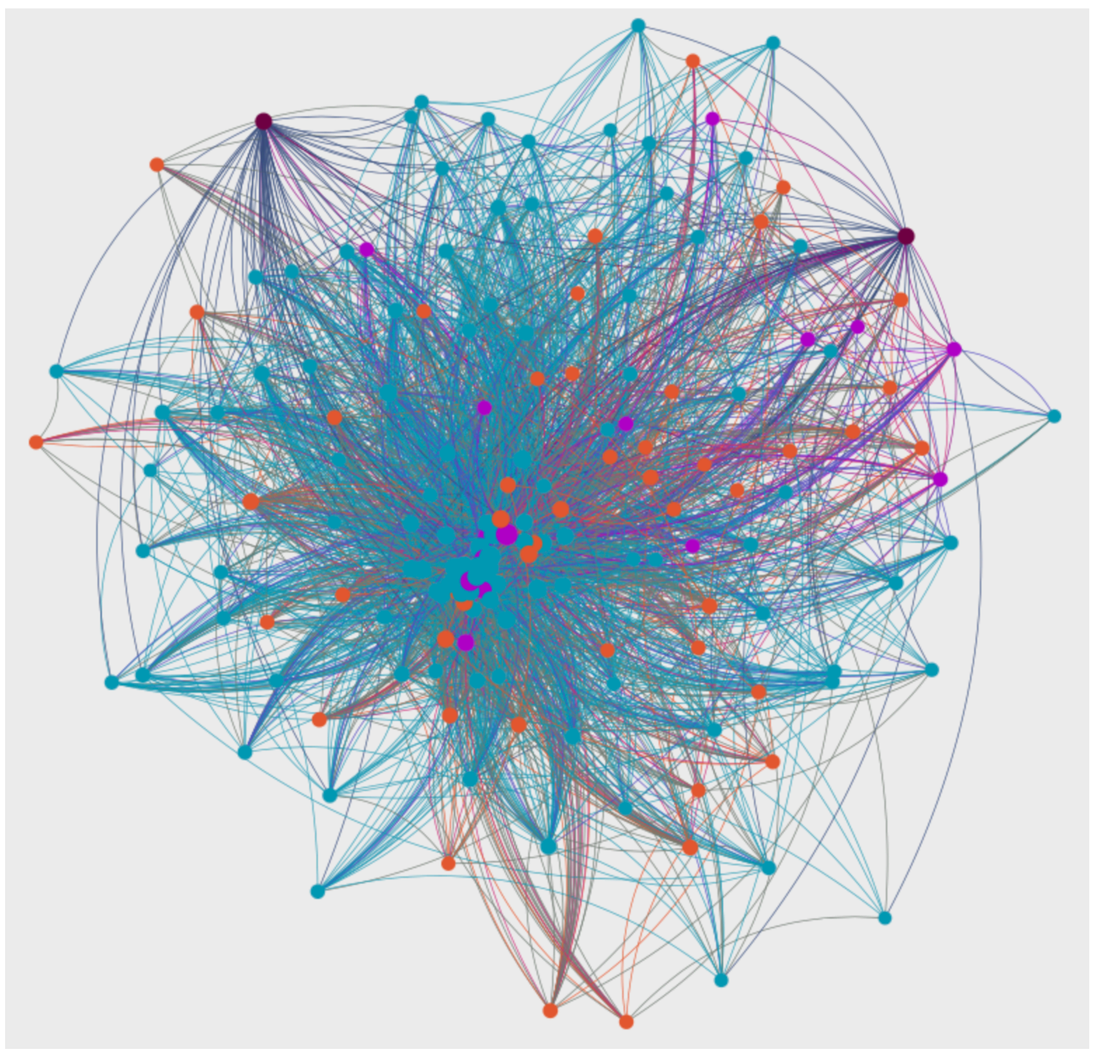
Background

- 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

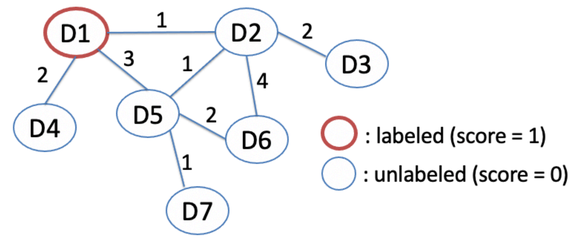


A network depicting the associations between phenotypes in our UKBB dataset – brown nodes represent source pre-eclamptic phenotypes, blue nodes represent associated diseases according to PheWAS summary data, orange nodes represent associated diseases according to UKBB case occurrences, and purple nodes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

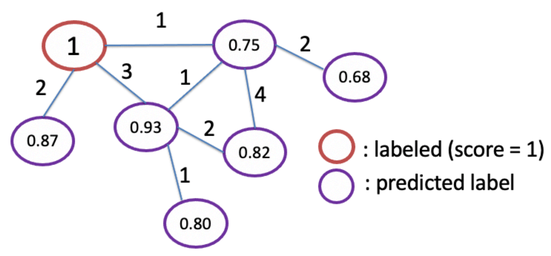
Graph-based Machine Learning Inference

- 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

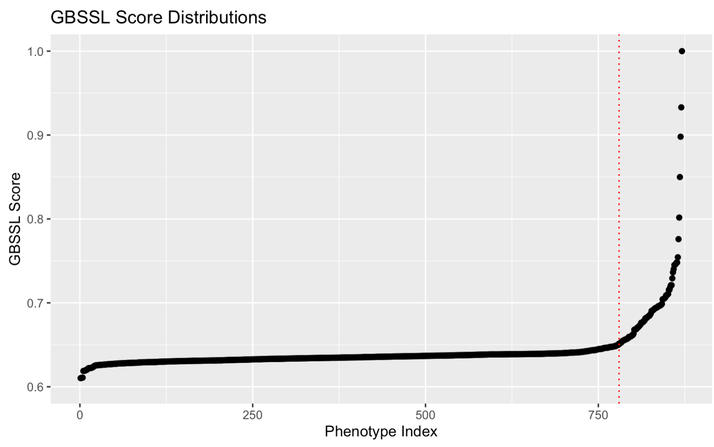


Optimal solution $\boldsymbol{f} = \{\boldsymbol{I} + \mu L\}^{-1}\boldsymbol{y}$,

where $L = D - W$, $D = diag(d_i)$, and $d_i = \sum_j w_{ij}$



A depiction of the Graph-based Semi-supervised Learning inference process. Scores are calculated for unlabeled nodes according to edge weights and network topology



GBSSL yields a distribution of scores representing strength of association with pre-eclampsia. The elbow in the curve can be used as a benchmark to establish which phenotypes are associated according to our method

- In order to evaluate the accuracy of labels, we considered clinical data alone from the UKBB electronic health records
- We compared occurrences of diseases in pre-eclamptic patients to occurrences in controls using the Chi-squared test

Results

Source Phenotype	Neighboring Phenotype	Number of Shared SNPs
Preeclampsia and eclampsia	Acquired absence of breast	15
Gestational Hypertension	Multiple Sclerosis	8
Gestational Hypertension	Scar conditions and fibrosis of skin	6
Gestational Hypertension	Pathologic fracture	5
Gestational Hypertension	Pelvic Inflammatory Disease	4
Preeclampsia and eclampsia	Diseases of sebaceous glands	4
Preeclampsia and eclampsia	Sebaceous cyst	4
Preeclampsia and eclampsia	Gestational Hypertension	3
Gestational Hypertension	Other hypertrophic and atrophic skin conditions	3
Preeclampsia and eclampsia	Hypothyroidism	3

Phenotype Name	GBSSL Score
Pathologic fracture	1.00
Scar conditions and fibrosis of skin	0.933
Acquired absence of breast	0.898
Other hypertrophic and atrophic conditions of skin	0.850
Abnormal serum enzyme levels	0.802
Contracture of joint	0.776
Abnormal findings on examination of biliary tract	0.754
Other complications of pregnancy NEC	0.748
Placenta previa and abrupto placenta	0.748
Anterior horn cell disease	0.746

Phenotype name	FDR adjusted P-value
Fetal abnormality affecting management of mother	0
Other complications of pregnancy NEC	9.72E-231
Complications of labor and delivery NEC	2.43E-174
Early onset of delivery	8.56E-161
Miscarriage; stillbirth	6.84E-159
Obstetrical/birth trauma	1.75E-157
Fetal distress and abnormal forces of labor	1.49E-98
Placenta previa and abrupto placenta	3.21E-90
Essential hypertension	3.45E-89
Malposition and malpresentation of fetus or obstruction	1.44E-71

Top 10 disease connections with pre-eclampsia in our DDN according to edge weight; top 10 disease connections according to GBSSL; top 10 disease connections according to chi-squared test

- 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. placenta previa, 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

Kim Lab
https://www.biomedinfolab.com

Vivek Sriram
viveksrm@pennmedicine.upenn.edu