Network analysis and visualization for electronic health records data

By

Nicholas James Strayer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

December 12, 2020

Nashville, Tennessee

Approved:

Mathew Shotwell, Ph.D.

Yaomin Xu, Ph.D.

Simon Vandekar, Ph.D.

Douglas Ruderfer, Ph.D.

This thesis is dedicated to my kind and brilliant wife, Sarah
and our cat, Flumpert.

ACKNOWLEDGEMENTS

Getting to this point in my research has been done in the company of many brilliant and inspiring people. The names that follow are in no way a complete list but represent those my faulty memory can conjure.

First and most directly, my advisor Yaomin Xu has helped shape me from a general bundle of interests and skills into something that may well be called a biostatistician and data scientist. Every member of the TBI-Lab has given me more than my due of inspiration and ideas to work with by sitting through my droning lab meeting presentations and managing to stay awake enough to provide meaningful feedback. Much of my work on networks has been done in the company of Siwei Zhang, who has provided both statistical rigor but valuable questions.

Various funding sources have given me the freedom to pursue my research questions unencumbered: The Vanderbilt Graduate Fellowship, The NIH Big Data To Knowledge training grant, and the Vanderbilt department of Biostatistics Development Award.

The members of my committee Matt Shotwell, Simon Vandekar, and Doug Ruderfer, have done an amazing job helping me weave a series of projects into a cohesive narrative that constitutes a dissertation. Special thanks belong to Doug Ruderfer, who has provided invaluable career advice as both a committee member but also a collaborator.

I have had the pleasure of doing various internships and jobs since starting my research career. At these positions, I have met many brilliant people and learned more than I thought I could. Here I will mention a few who have made particularly relevant impacts on me. Katharina Reineke, who, acting as my advisor for a summer research experience for undergraduates, helped show me what the life of a graduate student is like: an experience that inspired me to pursue my Ph.D. Jeff Leek, who somehow saw promise in a graduate student who procrastinated studying for comprehensive exams enough to write some javascript to swipe papers. The work I did at the JHU data science lab helped show me that writing robust software is data science and inspired me to procrastinate even more while learning about how to do it better. At the New York Times, I had the pleasure of meeting a ton of exceptional individuals, of which only a few I have room to mention. Archie Tse saw fit to give a graduate student with a passing knowledge of D3 an opportunity to reach hundreds of thousands of readers.

Amanda Cox somehow managed to know the answer to every question I ever asked before I could finish asking it. Yaryna Serkez encouraged me to investigate the more esoteric stories and also accompanied me on explorations of the city. Tim Wallace, who's advice on both finishing a Ph.D. and the best burritos in the city, gave me food for thought and body.

While my time at my alma mater UVM is rapidly disappearing into the past, the lessons I learned from the professors there are still wonderfully fresh. Helga Schreckenberger taught me to expect more from myself in the first-year "pursuit of knowledge" seminar and our subsequent lunches on the top of Waterman Hall. James Bagrow, who's data science and visualization class, was the tipping point that sent me down this journey. Jason Stockwell took on the giant task of managing to train me to be useful in an aquatic ecology lab. Richard Single, who's statistical genetics class and our manuscript on asymmetric linkage disequilibrium, showed me the tremendous power of merging statistics and data science with meaningful biomedical data.

One great set of advice I received once was to surround myself with people smarter than me, and I had no problem finding those people at Vanderbilt. Writing a blog with Lucy D'Agostino McGowan provided a series of inspiration every time she wrote 15 meaningful articles in a week, and I had written a typo-riddled intro to a single one. Alex Sundermann helped me see that superheroes exist, and they are brilliant epidemiologists who also happen to get MDs and beat me at Settlers of Catan.

Last - but in no way least - I have to thank my wife, Sarah. We met at the very start of graduate school, and ever since, I have been the lucky recipient of her incredible wisdom and insight. She's put up with me asking naive questions about genetics and taken the time to explain things well enough that even I can understand.

TABLE OF CONTENTS

Page

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF ABBREVIATIONS 0

Chapter

1 Introduction 1

1.1 Outline 1
1.2 Electronic health records 1
1.2.1 Billing codes 2
1.2.2 Phecodes 2
1.2.3 Biobanks 2
1.2.4 Is it worth it? 3
1.3 Network science 3
1.3.1 Types of networks 4
1.3.2 Basic networks primer 5
1.3.3 One-mode projections 7
1.3.4 Computing with networks 8

2 PheWAS-ME: A web-app for interactive exploration of multimorbidity patterns in PheWAS 10

2.1 Summary 10
2.2 Introduction 10
2.3 Implementation 11
2.4 Example usage 13

2.4.1 Exploration process . . . 14
2.4.1.1 Loading app . . . 14
2.4.1.2 Broadening selection . . . 15
2.4.1.3 Filtering to minor allele carriers . . . 15
2.4.1.4 Removing noisey phenotypes . . . 15
2.4.1.5 Using upset filtering to probe comorbidity frequencies 15
2.4.1.6 Removing low-sample-size phenotypes . . . 17
2.4.1.7 Investigating phenotype clusters . . . 17
2.4.1.8 Exploring cluster bridging phenotypes . . . 18
2.4.1.9 Exporting results . . . 18
2.4.2 Application State Links . . . 18
2.5 Discussion . . . 21

3 Interactive network-based clustering and investigation of association matrices with associationSubgraphs . . . 22

3.1 Summary . . . 22
3.2 Introduction . . . 22
3.3 Methods/Implementation . . . 24
3.3.1 Algorithm . . . 24
3.3.2 Visualization . . . 26
3.3.3 Choosing “optimal” threshold . . . 27
3.3.4 Simulating an association network . . . 28
3.3.5 Determining the performance of each step in the visualization 28
3.4 Results . . . 29
3.4.1 Performance of giant-component cutoff on simulated data . . 29
3.4.1.1 Visualizing the thresholds . . . 29
3.4.1.2 Scaling up simulations . . . 30
3.5 Visualizing comorbidity associations . . . 32
3.6 Discussion . . . 32

4 Investigating Phenome-wide comorbidity landscapes across two large-scale EHR systems . . . 34

4.1 Summary . . . 34
4.2 Introduction . . . 34
4.3 Methods . . . 37
4.3.1 Individual-level data . . . 37
4.3.2 GLM-based comorbidity strength . . . 37
4.3.3 Comorbidity patterns: . . . 38
4.3.4 Comorbidity similarity . . . 39
4.3.5 Phecode comorbidity conservation . . . 39
4.3.6 System-level comorbidity conservation . . . 40

4.3.7 Phenotype centrality . . . 40
4.3.8 Combined comorbidity network . . . 41
4.3.9 Subgraph neighborhoods . . . 41
4.4 Results . . . 42
4.4.1 Patient populations . . . 42
4.4.1.1 Demographics . . . 42
4.4.1.2 Phecode prevalences . . . 44
4.4.2 Direct comorbidity results . . . 44
4.4.2.1 How conserved are comorbidity patterns across the two systems? . . . 44
4.4.2.2 Phecode centrality differences . . . 45
4.4.3 A combined comorbidity network . . . 49
4.4.3.1 UMAP projections . . . 52
4.4.3.2 Neighborhood clustering . . . 52
4.4.4 App to explore phecode comorbidity . . . 53
4.4.5 Investigating 295.1 - Schizophrenia . . . 53
4.4.5.1 Differences between systems . . . 54
4.4.5.2 Combined comorbidity network . . . 54
4.4.5.3 Contrasting comorbidity with Polygenic Risk Score . . . 56
4.5 Discussion . . . 56

5 Conclusion . . . 60

6 Appendix: Simulating comorbidity associations . . . 62

6.1 Simulation procedure . . . 62
6.1.1 Generating patient data . . . 62
6.1.2 Calculating the association network . . . 63
6.2 Results . . . 63
6.2.1 Association distributions . . . 63
6.2.2 Association correlations . . . 64
6.3 Discussion . . . 64

REFERENCES . . . 66

LIST OF TABLES

Table		Page
4.1	Age distribution differences between two systems	43
4.2	Most extreme conservation levels across phenome.	47
4.3	Largest differences in centrality across systems	50
4.4	Top differences in 295.1 comorbidity between systems	55
4.5	Most comorbid phecodes with 295.1 in the combined comorbidity network	56
4.6	Comorbidity of 295.1 to phecodes significantly associated with PRS	58
6.1	Statistics for comorbidity z values for simulated association pairs	63

LIST OF FIGURES

Figure		Page
2.1	Screenshot of PheWAS-ME running on SNP rs200445019.	12
2.2	PheWAS-ME usage example step 1	14
2.3	PheWAS-ME usage example step 2	15
2.4	PheWAS-ME usage example step 3	16
2.5	PheWAS-ME usage example step 4	16
2.6	PheWAS-ME usage example step 5	17
2.7	PheWAS-ME usage example step 6	18
2.8	PheWAS-ME usage example step 7	19
2.9	PheWAS-ME usage example step 8	19
2.10	PheWAS-ME usage example step 9	20
3.1	A classic hairball network visualization	23
3.2	Find neighborhood subgraph algorithm	25
3.3	AssociationSubgraphs screenshot	26
3.4	Normalized Mutual Information of associationSubgraph visualization to true cluster structure	29
3.5	Local and global giant-component cutoff thresholds	30
3.6	Example of optimal cutoff with simulated network	31
3.7	Screenshot of optimal cutoff for simulated network	31

3.8 Relative performance of optimal cutoff suggestions in simulation . . . 32

4.1 Comorbidity pattern explainer 39

4.2 Phecode similarity explainer 39

4.3 Comorbidity conservation explainer 40

4.4 Eigen-centrality explained 41

4.5 AssociationSubgraphs results for comorbidity network 42

4.6 Demographic statistics for both subset of populations for both systems 43

4.7 Prevalences of phecodes across both systems 44

4.8 All comorbidity strengths for both systems 45

4.9 Conservation of comorbidity patterns across phenome 46

4.10 Comorbidity patterns of least and most conserved phecodes 47

4.11 Phecode conservation patterns by category. 48

4.12 Phecode centralities across systems 49

4.13 Distributions of centrality differences by category. 51

4.14 Combined comorbidity distribution 52

4.15 UMAP embedding of combined comorbidity network 53

4.16 Comorbidity patterns of 295.1 across systems 54

4.17 Top comorbidities for 295.1 in combined network 55

4.18 Comorbidity to PRS score for 295.1 57

6.1 Distribution of simulated comorbidities 64

6.2 Heatmap of comorbidity correlations of simulated and real networks . 65

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Outline

This dissertation takes a network view of Electronic Health Records (EHR) using traditional statistical methods and network science to build new tools to discover and explore complex phenotype associations in noisy data. In chapter 2, we demonstrate how the injection of network statistics and visualization into traditional methods can empower the analyst to check model assumptions and explore the underlying data to find patterns previously complicated to uncover. Next, in chapter 3, we show how network algorithms and interactive visualization methods can be combined to construct a visualization tool for exploring extremely high-dimensional association structures, such as that of the clinical phenome. Last, in chapter 4, we demonstrate the utility and value of network statistics by applying custom-built but simple metrics to explore and quantify the similarity of two comorbidity networks generated from separate EHRs.

Before we start, we will go over the basics of EHRs and network science to provide the reader with the necessary background knowledge to utilize the later chapters' contributions.

1.2 Electronic health records

EHRs, in the simplest terms, are computer-stored patient records. As computers and computer storage become ever cheaper and more ubiquitous, hospitals have shifted from storing patient data on paper charts to using EHRs. The full benefits and implementations of EHRs are beyond this dissertation's scope, but brief examples include the ability for rapid retrieval of a patient's data from any computer within the hospital's network, and the transfer of that information efficiently to new care providers.

EHRs are a general concept and contain many types of data. Examples include everything from hand-written notes to high-resolution radiology scans. Here, we focus on a small subset known as "billing codes."

1.2.1 Billing codes

In the form of ICD9 (World Health Organization, 1978) and the newer ICD10 (World Health Organization, 2004) codes, billing codes are an internationally recognized standard list of codes used to characterize a patient's stay in the hospital for billing purposes. As the range of possible conditions is vast, so are the number of codes present. ICD9 codes contain around 13 thousand different codes, and the newer ICD10 standard has around 68 thousand. While their extreme dimensionality makes the use of ICD codes in statistical modeling difficult, they are further hindered by being designed for billing and not research. This difference may seem subtle but means any models and inferences made from ICD data will have numerous and complicated to discern biases. For further information on the challenges of analysis with ICD codes, we refer the reader to chapter 4 of this dissertation and (Yadav et al., 2018).

1.2.2 Phecodes

In an attempt to alleviate some of the issues associated with ICD code usage in our analyses, ICD9 and ICD10 codes present in a patient's records are mapped to the lower-dimensional phenotype classifications called "phecodes." Phecodes were developed in parallel with the Phenome-Wide Association Study (Denny et al., 2010) (PheWAS) to translate the billing oriented ICD codes to a more research appropriate mapping of around 1.8 thousand clinically relevant phenotypes. Phecodes benefit from interpretable descriptions and high-level category groupings. These clean and easy to understand groupings allow for "sanity-checks" of results and open up interpretation of results to non-physician scientists.

1.2.3 Biobanks

In addition to the use of billing code data, some parts of this dissertation use data from "biobanks." Biobanks are repositories of biomarker data from patients in a hospital system or systems. These biomarkers can be a large variety of things, from raw serum up to single-cell RNA sequencing data. There are many exciting applications of the merger of biobank and billing-code derived phenotypes; however, here we touch only on the aforementioned PheWAS studies, which seek to discover associations between a given biomarker of interest and each distinct phenotype present, or the user's "Phenome." These studies, like their predecessors, the Genome-Wide

association study (Hindorff et al., 2009) (GWAS), typically use Single Nucleotide Polymorphisms (SNPs) as their biomarkers of interest.

1.2.4 Is it worth it?

With billing-code derived phenotypes subject to so many potential confounding and noise-introducing biases, it is reasonable to question why any effort should be made to build models around them. Skepticism is valid and valuable in these cases, but ultimately the potential upsides make the efforts worth it. With the clinical care industry worth $3.6 trillion as of 2018 (hea, 2020), the purely monetary benefits of successful translational research done on billing code data are vast.

Even with the risks mentioned above, we have seen great successes from billing-code based work (Ritchie et al., 2010). For example, the use of PheWAS studies to find drug repurposing targets has successfully identified alternative targets for already approved drugs (Werfel et al., 2020). This type of work not only saves money but, more importantly, speeds up the availability of potentially life-saving medications. Further, in chapter 4 of this dissertation, we show that the phenome's topological structure, as represented by phecodes, is surprisingly robust.

1.3 Network science

> "A network is a simplified representation that reduces a system to an abstract structure of topology, capturing only the basics of connection patterns and little else… This has some disadvantages but it has advantages as well." - Newman, Networks chapter 1 (Newman, 2018)

While network science is a relatively new discipline, its roots trace back to work done by Leonard Euler in 1741 (Euler, 1741). The likely first problem with a formal network-based solution is known as "The Seven Bridges of Königsberg." The goal of the seven bridges problem was to build a path through the Prussian city of Königsberg that crossed each of its seven bridges just once. By breaking the problem down to one of a network, representing each landmass (either side of a river and an island within the river) as nodes and the bridges connecting them as edges, Euler was able to show with mathematical formality that there was no solution to the problem. Euler's seven bridges result marked the first widely recorded use of networks as an abstract topological representation of a system, but it was not the last.

1.3.1 Types of networks

The networks used in network science can come from any system that can be abstracted to a series of nodes and connections between those nodes; however, a few common areas have emerged as sources of network data.

One of the canonical examples of a network is a social network. Here, nodes are individuals, and connections are interactions in some form, such as friendship or coworking status. One of the first examples of a social network is the Southern Women's Cohort network (Davis et al., 2009), which tracked 18 women and the social events they attended in 1939. A common task with social networks is determining groups (or clusters) of individuals, such as political party separation (Minot et al., 2020).

Another example of network data commonly seen comes from technological networks. These typically directly represent the data connections between computers or other circuits. The first examples come from telegraph networks (Müller, 2016), and the much larger and more recent examples come from the internet (LYON and B, 2005). Here the questions posed tend to be more algorithmic, for example, how to efficiently route network packets between a server and a client (Ng et al., 2007).

As a method of connecting units of information (e.g., Wikipedia pages), the internet also represents another common type of network: Information networks. These networks - predictably - represent the flow of information, either abstractly or mathematically defined (e.g. entropy) between their nodes. Citation networks are another classic example of information networks. Whereby the nodes are researchers, and edges represent the co-authoring of a paper together. Potential questions asked include what disciplines tend to collaborate the most often (Newman, 2001).

Biochemical or biological networks represent one of the newer applications of network science. Here the networks can represent a large number of related biological processes. Examples range from the highly algorithmic: how to assemble a genome (Compeau et al., 2011; Pop, 2009); to information-based: genes co-expressing under certain conditions (Kovács et al., 2019).

Of all the networks mentioned, the work in this dissertation most closely relates to these biological networks. The problems posed mirror many of those seen in gene-regulation networks. What Phenotypes tend to occur together or are highly "comorbid?" Throughout the following chapters, we will draw inspiration from previous works on all network types to build new methods to explore and understand

Phenome-based networks.

1.3.2 Basic networks primer

Throughout this dissertation, a large amount of network science jargon is used. This section aims to provide both a reference point for those terms and a very brief overview of basic network science concepts. A more thorough introduction of these terms and concepts is available in the textbooks Networks by Newman (Newman, 2018), and Statistical Analysis of Network Data by Kolaczyk (Kolaczyk and Csárdi, 2020).

Node/Vertex: The node is the most basic unit of a network. It can represent almost anything from nations to individuals, down to the base pairs of those individuals genomes. Another name for "node" is "vertex." In this dissertation, the two terms are interchangeable. Often "node" is used in the description of real-world networks, and "vertex" is used when discussing abstract mathematical properties.

Bi/Polypartite networks The nodes in a network need not always be the same type of thing (e.g., all nodes are individuals). Many systems are more naturally described with multiple node "types," or as "polypartite" networks. The most common example of this is "bipartite" networks or networks with two node types. Examples of bipartite networks include states (or provinces) and their nations, students and the schools they attend, or genes and the samples in which those genes were expressed.

The only strict rule with polypartite networks is that two nodes of the same type can not share an edge. For instance, in the students-to-schools example, a student can be connected to a school as the edges represent attendance, but a student cannot "attend" another student, and neither can a school attend another school. These restrictions manifest themselves in the formation of the abstraction and are usually not challenging to accommodate.

While bipartite networks are the most common type after unipartite networks, there is no reason why the number of types is limited to two. For example, a network with nodes representing patients, genes, and phenotypes is a tripartite network where an edge between a patient and a gene represents the patient having a mutation on that gene, and an edge between a patient and a phenotype represents that patient having the phenotype in their medical records.

Edge/Link: An edge represents some form of connection or interaction between two nodes. Again, this is very general and can represent a vast range of "interactions."

For instance, the trading of goods between two nations, two individuals knowing each other, or two codons occurring side-by-side on the genome. Like "node" and "vertex," an edge is sometimes called a "link;" again, here we use the two terms interchangeably.

Weighted edge: Often an edge is annotated with information such as the strength of the connection. While these annotations can take any form, the most common is a scalar value that makes the edge "weighted." These values could be the monetary value of trades between two nations, how long two individuals have known each other, or how frequently two codons occur next to each other. While not strictly necessary, it is conventional for the weights to be exclusively positive values. Positive edge weights result in some convenient mathematical properties (see entry for "adjacency matrix"). As mentioned, this is not a hard-and-fast rule, and in chapter 4, we deal heavily with negative edge weights.

Directed edge: In addition to having weights, edges can also have direction. For instance: a nation's exports and imports, one individual may consider the other a "friend" but not necessarily the reverse, or asymmetric conditional probabilities of two codons co-occurring (Single et al., 2016; Thomson and Single, 2014). While directed edges can represent many exciting systems, this dissertation will primarily focus on un-directed networks.

Neighbors: A node's "neighbors" are all of the other nodes with which they share an edge, e.g., all the nations that trade with the United States, all of a student's friends at school, or all the codons that co-occur with AGT.

Degree: The degree of a node is the sum of all its edges its adjacent edges. In the case of the typical binary connections of an unweighted network, this is simply the number of neighbors the node has. For weighted networks, the degree is typically reported as the sum of the weights of all a node's adjacent edges.

Graph: It is common - especially in algorithmic contexts - for a network to be referred to as a "graph." In this dissertation, we will treat "graph" and "network" interchangeably.

Subgraph: A subset of nodes and edges within a larger graph or network. In mathematical terms: a graph H is a subgraph of G if $V_H \subseteq V_G$ and $E_H \subseteq E_G$, where V_G is the set of all **v**ertices in H, and E_G is the set of all **e**dges in H.

Isolated subgraphs/Components: Often, within a network as a whole, some subgraphs are entirely isolated from the rest of the network. Isolated here meaning there is no way to step along a series of edges between nodes to reach the rest of

the network. In the context of landmasses connect by bridges in the seven bridges example (Euler, 1741), there is no way to get from landmass A to landmass B by crossing bridges. The most common term used to describe these isolated subsets of nodes is "component," however, in this dissertation, we will use the term "isolated subgraph" to avoid confusion with other statistical concepts used for high-dimensional data such as Principle Components Analysis.

Adjacency matrix: One popular way of representing the edges of a network is via an "adjacency matrix." In a network with n nodes, this is an $n \times n$ symmetric matrix with the edge weight between nodes i and j encoded in the value of the i, jth cell. The diagonal elements are taken to be 0. The adjacency matrix is used heavily in mathematical representations and theory for networks such as calculating the eigen-centrality of nodes. Eigen-centrality ranks nodes "importance" within the network by taking the n elements of the eigenvector with the largest eigenvalue to be the node's importance value. All the entries in the adjacency matrix being positive are necessary for at least one eigenvector to itself be all positive, as stated by the *Perron-Frobenius theorem.* For more details on eigen-centrality, see chapter 4 of this dissertation and chapter 7 of (Newman, 2018).

1.3.3 One-mode projections

Mathematical simplicity and an abundance of example datasets has resulted in most network analysis methods being designed for unipartite networks. This unipartite bias means that it is common for polypartite networks to be first collapsed or "projected" to a unipartite or "one-mode" version. This new collapsed network can then have the unipartite algorithms run on it.

The most straightforward format of this collapsing is co-occurrence. In the students-schools bipartite network example, this would mean connecting students if they both share a connection to (aka attended) the same school. While the methods used to collapse networks can get much more complicated (see 4.3.2 for an example), there is always a loss of information about the system that occurs when simplifying the structure (Larremore et al., 2014). This loss of information means that collapsing should only be done when appropriate. However, as we will see in chapter 4, sometimes this information loss is a feature that can be used to preserve the anonymity of nodes within the original un-collapsed network.

1.3.4 Computing with networks

Now that datasets are large and compute so cheap, practically all network analysis is done on a computer. While much of the math underpinning network theory is built around manageable mathematical constructs like the adjacency matrix, these representations can be highly inefficient when translated onto a computer. As a result, a large body of data structure and algorithms development has been devoted to representing and operating on networks. Here we will provide a very brief overview of common network representations and algorithms. We point readers to chapter 8 of (Newman, 2018) and chapter 4 of (Sedgewick and Wayne, 2011) for more complete coverage of this topic.

Storing networks: The most common method of representing a network on a computer is through objects representing nodes that contain an array of references to every one of that node's neighbors, similar to a linked-list. This representation has the benefit of providing fast access to the node's neighbors, allowing rapid traversal of the network. The ability to rapidly "walk" along the edges of a network is a fundamental element of network algorithms.

Traversal algorithms: While there is a large number of computing algorithms that operate on networks, almost all of these are variations on two similar algorithms for traversing the network: **breadth-first search (BFS)** and **depth-first search (DFS).** By using the linked representation of a network, these algorithms and their variations avoid exhaustive searching, meaning that they often run in sub-linear time, typically $O(\log n)$.

Both breadth- and depth-first search are relatively simple and follow a similar template:

1. Start with a given (often random) node in the network
2. Add every neighbor of the current node to a list of "nodes to explore"
3. Starting at the top of the nodes-to-explore list, repeat the steps **1** and **2**
    - For BFS, add all neighbors to the *bottom* of nodes-to-explore
    - For DFS, add all neighbors to the *top* of nodes-to-explore
4. Once all n nodes have been explored, or the desired node is found, the algorithm finishes

The difference of which end of the nodes-to-explore list the neighbors are added determines if the algorithm searches wide and shallow (i.e., with "breadth"), or if

it exhaustively explores a given edge path before resetting (i.e., with depth). Both algorithms are used in similar scenarios, such as finding the shortest paths through mazes (MOORE and F, 1959; Sedgewick and Wayne, 2011). The choice between which algorithm to use typically depends on properties of the network being explored, but for non-tree-based networks - like those considered in this dissertation - DFS typically is the variant of choice (Sedgewick and Wayne, 2011; Newman, 2018).

CHAPTER 2

PHEWAS-ME: A WEB-APP FOR INTERACTIVE EXPLORATION OF MULTIMORBIDITY PATTERNS IN PHEWAS

2.1 Summary

Electronic health records linked with a DNA biobank provide unprecedented opportunities for biomedical research in precision medicine. The Phenome-wide association study is a widely-used technique for the evaluation of relationships between genetic variants and a large collection of clinical phenotypes recorded in EHRs. PheWAS analyses are typically presented as static tables and charts of summary statistics obtained from statistical tests of association between a genetic variant and individual phenotypes. Comorbidities are common and typically lead to complex, multivariate gene-disease association signals that are challenging to interpret. Discovering and interrogating multimorbidity patterns and their influence in PheWAS is difficult and time-consuming. We present PheWAS-ME: an interactive dashboard to visualize individual-level genotype and phenotype data side-by-side with PheWAS analysis results, allowing researchers to explore multimorbidity patterns and their associations with a genetic variant of interest. We expect this application to enrich PheWAS analyses by illuminating clinical multimorbidity patterns present in the data.

Availability: A demo PheWAS-ME application is publicly available at https://prod.tbilab.org/phewas_me/. Sample datasets are provided for exploration with the option to upload custom PheWAS results and corresponding individual-level data. Online versions of the appendices are available at https://prod.tbilab.org/phewas_me_info/. The source code is available as an R package on GitHub (https://github.com/tbilab/multimorbidity_explorer).

2.2 Introduction

Large-scale biobanks combined with electronic health records are increasingly available for clinical and translational research around the world (Chen et al., 2011; Cho et al., 2012; Gaziano et al., 2016; All of Us Research Program Investigators et al., 2019; McCarty et al., 2011; Sudlow et al., 2015,). These data platforms typically

provide subject-level information on a wide range of biomarkers along with detailed phenotype data and provide a highly anticipated paradigm shift for clinical and translational research in the era of precision medicine. The Phenome Wide Association Study is a statistical method to find associations across phenomes in the EHR with a given biomarker (e.g. SNPs). PheWAS quantifies associations between single SNP-phenotype pairs, which are blind to complex correlation structures present in phenotypes. When multiple phenotypes show a strong association with a genetic variant, researchers rely on domain expertise and more extensive interrogation of the data to determine potential causes. These include driver phenotypes (e.g., patients with a common disease taking a drug and then experiencing a common drug side effect), phenotype hierarchy, related diseases with an overlapping set of patients, or merely people with multiple diseases. Here we present PheWAS Multimorbidity Explorer (PheWAS-ME), a web application built using the programing language R and the Shiny library (Chang et al., 2020). PheWAS-ME allows researchers to interact with PheWAS results alongside the individual-level phenotype and genotype data that generated them. By visualizing individual-level data along with statistical results, the application provides a rich and explorable view into the patterns and relationships between phenotypes and the genotype being investigated. The interactive nature of the tool lets users enhance their interrogation of comorbidity patterns by delving into areas of interest on the phenome, such as a disease category, with custom visualizations. See Appendix B for a demonstration of the use of PheWAS-ME to parse the results of a PheWAS analysis to find novel phenotype associations.

2.3 Implementation

Data needed to run PheWAS-ME are a standard PheWAS result table and the corresponding individual-level data. These results can be supplied to the app via a data loading screen or pre-loaded (see appendix C for full requirements). If desired, multiple comparisons correction can be performed on loaded data using either the Bonferroni (Dunn, 1961) or Benjamini-Hochberg (Benjamini and Hochberg, 1995) methods.

After data are loaded, the app directs to the main visualization and analysis interface - an interactive dashboard including four views: SNP information, an interactive PheWAS Manhattan plot, multimorbidity UpSet plot, and a subject-phenotype bipartite network plot.

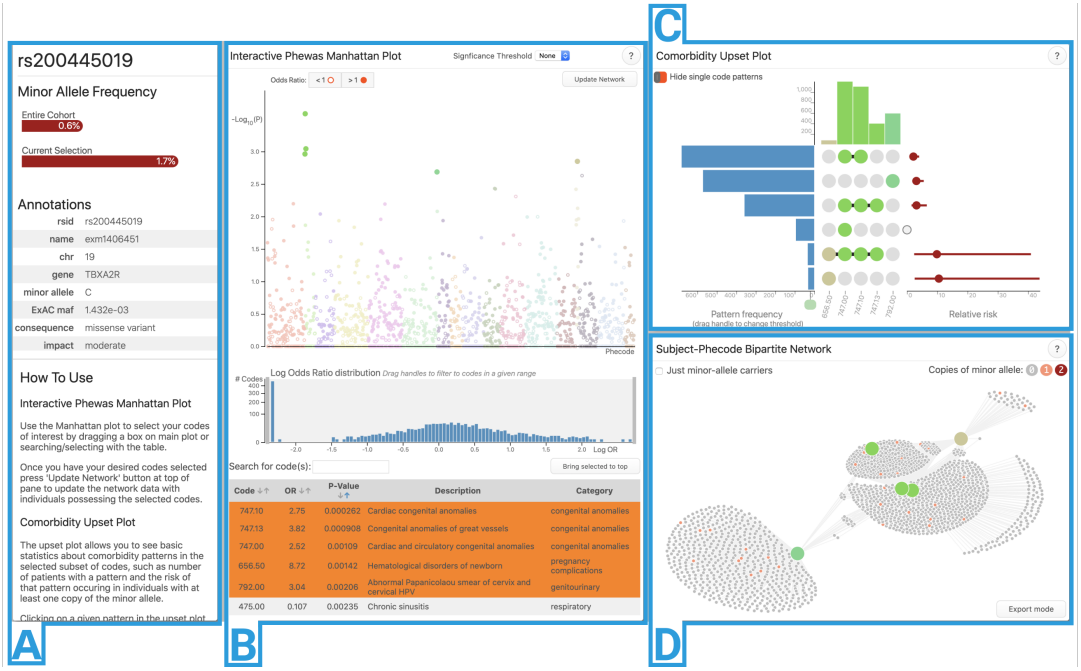

Figure 2.1: Screenshot of PheWAS-ME running on SNP rs200445019.

Application state: PheWAS-ME works by filtering down to a list of 'selected' phenotypes. When a set of phenotypes is selected, the individual-level data are subset to just subjects who had one or more of the selected phenotypes in their records. This allows users to easily discard uninteresting or noisy phenotypes and focus in on potentially meaningful patterns using criteria like strength of the statistical association or phenotype category.

SNP information panel: To provide context to the currently investigated SNP, the application provides a panel containing summary information (Figure 2.1A). Minor allele frequency in the provided subject population and the currently selected subset are shown as a bar chart. If the SNP of interest is present in an internal SNP annotation table sourced from dbSNP (Sherry et al., 2001) and VEP (McLaren et al., 2016), then additional information such as the minor allele, chromosome, and gene are provided.

Interactive PheWAS Manhattan plot: A manhattan plot (Figure 2.1B) is provided for the results of the PheWAS analysis (Denny et al., 2010). The x and y axis of this plot are phenotype diagnosis and statistical significance, respectively. Additional metadata from the supplied results table - such as name, description, and statistical results for a phenotype - are accessible by hovering over a phenotype's point in the plot. Phenotypes can be selected for individual-level-data inspection by any combination

of clicking, dragging a selection box, and searching in a table view below the plot. A histogram of the log-odds ratios for all phenotypes is provided and can be used to filter codes by ranges of association strength and direction.

Multimorbidity UpSet plot: Figure 2.1C is an UpSet plot (Lex et al., 2014). This plot shows the unique multimorbidity patterns seen in the individual-level data for the currently selected phenotypes as a matrix with columns as phenotypes and patterns (represented by filled phenotype columns) as rows. On the left side of the plot is a bar-chart displaying how many subjects had a multimorbidity pattern. To the right is a point estimate and 95% confidence interval of each pattern's relative risk of occurring given that the subject has the given genetic variant of interest (calculated using Fisher's Exact Test (Fisher, 1992)). When a pattern is selected, the subjects who have the pattern are highlighted in the network plot (Figure 2.1D). Hovering over a phenotype's column displays its name and description, supporting the quick examination of multimorbidity pattern membership. For more details on the upset plot we refer the reader to the original UpSet publication (Lex et al., 2014).

Subject-Phenotype Bipartite Network: Individual-level data are visualized directly as a bipartite network. Phenotypes are represented as larger nodes (colored to match their point in the manhattan and upset plots) and subjects are represented as smaller nodes (colored by their number of copies of the SNP minor allele). A link is drawn between subjects and phenotypes if a subject was diagnosed with a phenotype. A physics-based layout simulation (Bostock et al., 2011) is run in real-time as the data are filtered to position nodes with similar connection patterns close to each other. As the user investigates the network structure, phenotype nodes can be selected and isolated or removed from within the plot. An optional filtering mode limits the network to only subjects with one or more copies of the SNP's minor allele, allowing investigation of genetics-driven patterns. An "export mode" button lets the user download a high-resolution copy of the plot with optional phenotype labels for use in publications.

Greater detail of each section of PheWAS-ME is available via in-app help pages and the meToolkit package usage manual (see "Availability").

2.4 Example usage

Here we describe a sample use-case of the PheWAS-ME application for investigating the results of a PheWAS analysis and subject-level data corresponding to the SNP

Figure 2.2: State of application at initial load

rs200445019.

Published literature shows the gene TBXA2R is associated with advanced cardiovascular disease (Yi et al., 2019; Milanowski et al., 2017; Bauer et al., 2014; Schumacher et al., 1992). A recent paper (Werfel et al., 2020) demonstrates a novel association of SNP rs200445019 - a mutation within TBXA2R that is known to enhance a protein receptor regulating coagulation, blood pressure, and cardiovascular homeostasis - with cancer metastasis phenotypes. A finding validated using mouse models. Here we show how PheWAS-ME is capable of uncovering this new phenotype association by providing a flexible set of interaction workflows to explore PheWAS analysis results for rs200445019 and accompanying subject-level data.

The application used in this demonstration is available at https://prod.tbilab.org/phewas-me-rs200445019. The section 2.4.2 contains links to the various application states described in this walkthrough.

2.4.1 Exploration process

2.4.1.1 Loading app

PheWAS-ME starts in the default state: with the top 5 most significant codes visible.

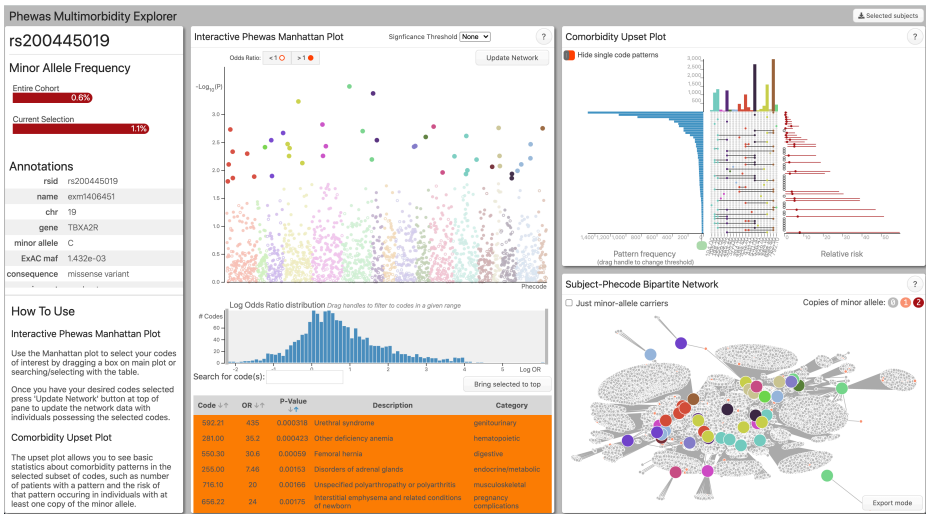

Figure 2.3: The region selection tool in the manhattan plot is used to draw a selection around the most significant phecodes.

2.4.1.2 Broadening selection

The P-Value threshold line and manhattan-plot region selection are used to expand the selected phecodes to those with a P-Value < 0.01 along with codes just below the threshold that appear separated from the background level of significance. (Corresponding to a P-Value < 0.015)

2.4.1.3 Filtering to minor allele carriers

The network panel is filtered to only minor-allele carriers to view the individual data in the context of the SNP. This filtering allows the network to reflect the genetic association network between phenotypes, rather than general comorbidity.

2.4.1.4 Removing noisey phenotypes

Using the PheWAS results table, nonspecific and general-purpose phenotypes - such as 303 (Psychogenic and somatoform disorders) and 782.30 (Edema) - are removed.

2.4.1.5 Using upset filtering to probe comorbidity frequencies

At this stage, the network plot appears to show relatively distinct clusters of comorbid phenotypes. For further insight into this separateness, the pattern frequency limiter is reduced to show all patterns with more than one subject. This expansion shows that only two phenotype comorbid pairs occur more than twice: 285.22 and 198.20

Figure 2.4: Using the checkbox "Just minor-allele carriers" in the network plot the subject-level data is reduced to only individuals who have one or more copies of the minor allele of interest.

Figure 2.5: With noisey phenotypes removed a separation into two clusters begins to appear in network plot.

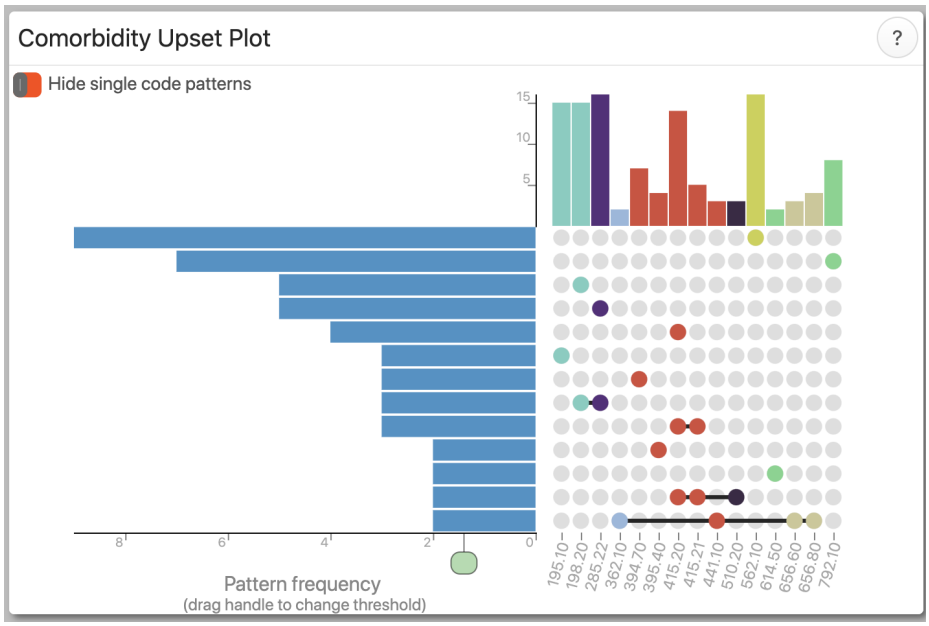

Figure 2.6: The upset plot panel complements the network plot by showing exact counts of comorbidity patterns in currently selected data.

("Anemia in neoplastic disease" and "Secondary malignancy of respiratory organs"), and 415.20-415.21 ("Chronic pulmonary heart disease" and "Primary pulmonary hypertension") and no comorbidity pairs cross multiple general phenotype topics.

2.4.1.6 Removing low-sample-size phenotypes

In order to avoid spurious separations in clusters due to limited sample size, all phecodes with less than three minor allele carrying subjects are found by mousing over phecodes in the network panel. These low-sample-size codes are removed from the currently selected phecodes using the "delete" option in the network selection context menu.

2.4.1.7 Investigating phenotype clusters

At this stage the noted two comorbid clusters in the subject-level graph persist. With the mouse-over tooltips the contents of these clusters can be investigated. This investigation reveals one cluster consisting mostly of circulatory system phenotypes (e.g., 394.70: "Disease of tricuspid valve" and 395.40: "Nonrheumatic pulmonary valve disorders") and the other consisting of a mixture of Neoplasm phenotypes (e.g., 195.1: "Malignant neoplasm, other" and 198.2: "Secondary malignancy of respiratory organs") and cancer treatment side-effects (e.g., 572.7: "Disturbance of salivary

Figure 2.7: With low-sample-size phecodes removed from the subject data, the distinction of two clusters remains.

secretion," 528.1: "Stomatitis and mucositis")

2.4.1.8 Exploring cluster bridging phenotypes

The relative separation of these patterns is reflected in the upset plot: only a single minor-allele-containing subject has phecodes in both cancer and circulatory phenotype clusters. By using the mouse-over tooltip in the network plot, the phenotype in the "cancer" cluster with a common subject is seen to be 562.1: "Diverticulosis": a highly age-linked phenotype. This separation suggests a potentially novel association between rs200445019 and cancer-related phenotypes. This association does not appear driven by correlation with the (previously known to be associated) coronary phenotypes.

2.4.1.9 Exporting results

The export functionality in the network panel is used to produce a vector-based output plot with all selected phecodes labeled for use in publication or to share with collaborators.

2.4.2 Application State Links

- **2.4.1.1:** https://prod.tbilab.org/phewas-me-rs200445019

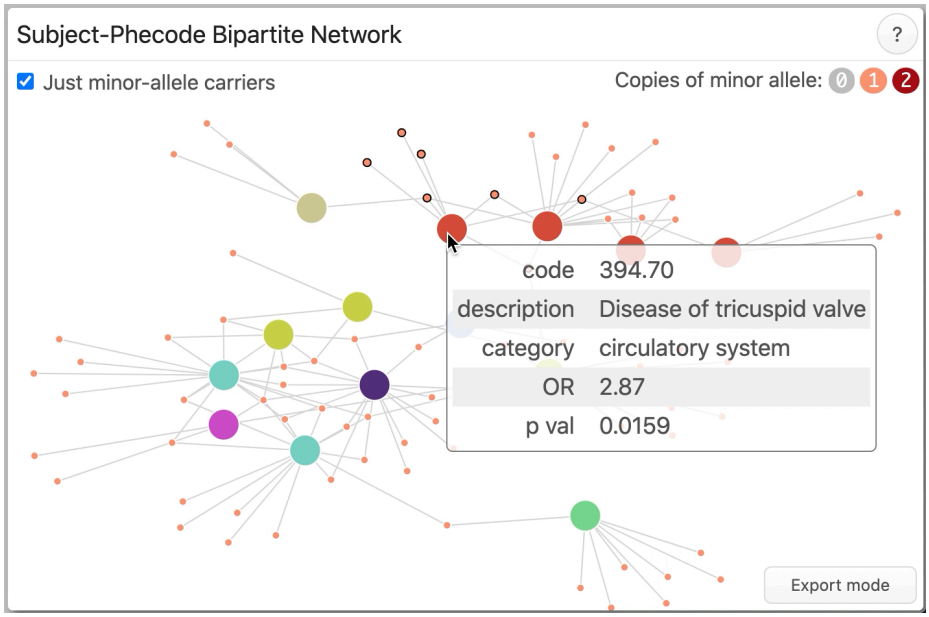

Figure 2.8: By mousing over phecode nodes in the network plot the two comorbidity clusters can be explored.

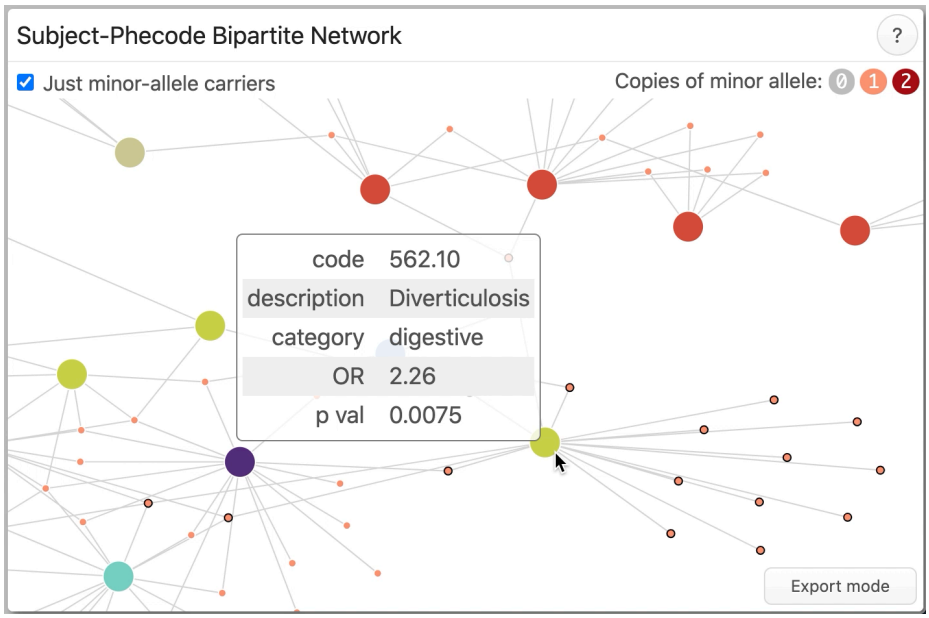

Figure 2.9: Mouseover information on the bridging phecode shows it is linked to coronary phenotypes by a single subject and is a highly age-related phenotype.

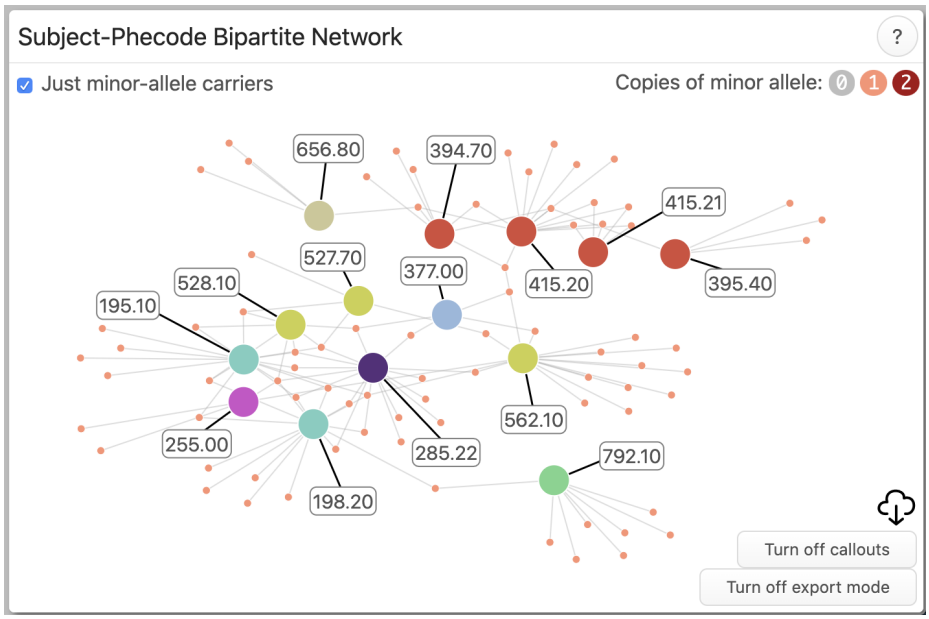

Figure 2.10: The final state of the network plot is put into export mode so phecode id callouts can be added for use outside of app.

- **2.4.1.2:** https://prod.tbilab.org/phewas-me-rs200445019/_w_086bb6d1/?rs200445019___41100_41520_93910_69410_75600_52641_52810_55030_59221_25500_25930_39540_44110_41521_39470_45300_52770_69510_56210_25512_28100_61450_28522_97600_98800_71610_30340_19510_79210_14900_73200_22700_19820_19400_65622_78230_37710_65660_35800_65680_36441_51020_36210_51000_37700
- **2.4.1.3:** https://prod.tbilab.org/phewas-me-rs200445019/_w_086bb6d1/?rs200445019___41100_41520_93910_69410_75600_52641_52810_55030_59221_25500_25930_39540_44110_41521_39470_45300_52770_69510_56210_25512_28100_61450_28522_97600_98800_71610_30340_19510_79210_14900_73200_22700_19820_19400_65622_78230_37710_65660_35800_65680_36441_51020_36210_51000_37700___ma_filtered
- **2.4.1.4:** https://prod.tbilab.org/phewas-me-rs200445019/_w_d23113f0/?rs200445019___41100_41520_93910_69410_75600_52810_55030_59221_25500_25930_39540_44110_41521_39470_45300_52770_69510_56210_25512_28100_61450_28522_19510_79210_14900_73200_22700_19820_19400_65622_65660_35800_65680_36441_51020_36210_37700___ma_filtered
- **2.4.1.6:** https://prod.tbilab.org/phewas-me-rs200445019/_w_086bb6d1/?rs200445019___41520_52810_25500_39540_41521_39470_52770_56210_

28522_79210_19820_65680_37700_19510___ma_filtered

2.5 Discussion

In this paper we have provided a brief introduction to the application PheWAS Multimorbidity Explorer. This application takes PheWAS results and individual-level data, and enables researchers interactively explore complex multimorbidity patterns in PheWAS analyses.

CHAPTER 3

INTERACTIVE NETWORK-BASED CLUSTERING AND INVESTIGATION OF ASSOCIATION MATRICES WITH ASSOCIATIONSUBGRAPHS

3.1 Summary

Making sense of association networks is vitally important to many areas of high-dimensional analysis. However, as the data-space dimensions grow, the number of association pairs increases in $O(n^2)$; this means traditional visualizations such as heatmaps quickly become too complicated to parse effectively. Here we present associationSubgraphs: a new interactive visualization method to quickly and intuitively explore high-dimensional association datasets using network science derived statistics and visualization.

Availability: An R package implementing both the algorithm and visualization components of associationSubgraphs is available at https://github.com/nstrayer/associationSubgraphs. Online documentation and usage examples are available at https://nickstrayer.me/associationsubgraphs/.

3.2 Introduction

Analysis of association or correlations between variables is an essential step in exploratory data analysis of high-dimensional datasets. In these scenarios, a dataset with many columns (, or measured variables,) without known or validated patterns of association between them is inspected using statistical and visualization methods to gain insight into how the variables interact. There are many different ways of establishing these interactions' strength, from as simple as the mutual occurrence of binary variables (Cha, 2007) to complex penalized regression models (Hallac et al., 2015; Tibshirani et al., 2005). Examples of areas where association analysis is used are gene regulatory networks (Gustafsson et al., 2005), analysis of single-cell sequencing data to determine cell differentiation (Chan et al., 2017), networks of comorbidity between diseases (Chen and Xu, 2014), topic modeling in natural language processing (Wang and Zhu, 2014), and many others.

The most common way of analyzing association patterns is to use heatmaps. In

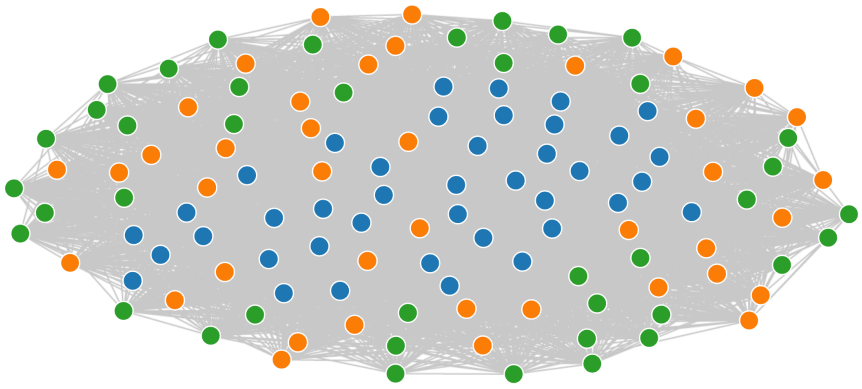

Figure 3.1: An example of the "hairball" problem in the visualization of dense networks. The density of overlaps and number of competing forces on the layout algorithm result in very poor inference of underlying structure. This is clearly seen here, where a true structure of three clusters of similar nodes are simulated but appear almost purely randomly placed in the visualization.

heatmaps, both rows and columns represent all present variables, and the color of the cells represent the strength of the association between the two variables. Once the number of variables gets large, the effectiveness of a heatmap rapidly decreases. The main issue present is the ordering of the rows and columns. What variables are placed next to each other can completely change inference made by the analyst (Bojko, 2009) and must be done with care. Typically this ordering is done with a non-trivial clustering algorithm (Metsalu and Vilo, 2015; Pryke et al., 2007), which injects model assumptions into the visualization that are not immediately clear to the analyst or later audience. Another issue with large heatmaps is the difficulty of discerning the identity of cells that fall far from the labeled axes (Bojko, 2009).

One way of alleviating the issues with heatmaps is to represent the association space as a network and visualize it using various network visualization tools. Unfortunately, beyond a few nodes, the visualization devolves into what is commonly referred to as a "hairball" network (figure 3.1) with little to no value in interpreting results.

An obvious way to deal with network representations' hairball issue is to do edge/association filtering and then visualize the reduced network. Methods used to do this filtering typically involve parametric tests that do not account for the network-structure (Benjamini and Yekutieli, 2001) or contain a large number of assumptions (Hallac

et al., 2015). Non-parametric methods typically based on permutations also exist, but due to the $O(n^2)$ complexity inherent in association datasets, they are computationally infeasible for large datasets (Harris and Drton, 2013). For more information on these methods, we point the reader to the book (Kolaczyk and Csárdi, 2020).

A classic model in network science is the "random graph" (Solomonoff and Rapoport, 1951, Erdös and Rényi (1959)) (often called the Erdos-Renyi graph). In random graphs, nodes are randomly connected by a given number of edges, i.e., without any specific node preference. An emergent property of these "random" graphs are "components" or "isolated subgraphs" (Newman, 2018, chapter 10) (In this paper, we will not use "component" to avoid confusion with principal-components). Isolated subgraphs are groups of nodes that are connected internally but not to any other nodes in the network. Percolation theory (Newman, 2018, chapter 15) is a subfield of network-science dedicated to understanding how the removal of edges within a network leads to the formation of these isolated subgraphs. By framing association analysis as a network problem, we can utilize the results of percolation theory to design an intuitive set of visualizations for exploring the structure of association networks based around the concept of adding and removing edges in order of the strength of association.

3.3 Methods/Implementation

3.3.1 Algorithm

The algorithm for computing associationSubgraphs at all given cutoffs is closely related to single-linkage clustering (Gower and Ross, 1969). However, it differs philosophically by viewing nodes that are yet to be merged with other nodes as "unclustered" (and thus unvisualized) rather than residing within their own cluster.

To calculate the set of subgraphs at every threshold, the algorithm starts by sorting edges/associations in descending order of strength. Then the nodes connected by the highest association strength are set as a "cluster." Next, the second-highest association strength is added. If either adjacent node is shared with the first cluster, then the non-shared node is added to the existing cluster. Otherwise, a new separate cluster containing the two adjacent nodes is created. This procedure is repeated for all association pairs. If both nodes for a pair already reside in separate clusters, they are merged into a new "super" cluster. After every edge is added, the current cluster state is exported. For further details see Algorithm 3.2.

```
Result: A list of subgraphs of nodes formed at every unique edge-weight threshold
Input: edges: array of all association strengths in network
       a_nodes: array of first corresponding node for strength in edges
       b_nodes: array of second corresponding node for strength in edges
/* Order of "first" and "second" corresponding nodes does not matter                */
Initialize: node_to_subgraph: node id → to subgraph id
            subgraph_to_members: subgraph id → array of ids of nodes belonging to the subgraph
            subgraphs_at_strength: strength of association → state of subgraph_to_members after
            each step

for i ← 1 to length(edges) do
    /* Gather node ids and subgraphs (if they exist) for those nodes                */
    node_a ← a_nodes [i];
    node_b ← b_nodes [i];
    node_a_subgraph ← node_to_subgraph(node_a);
    node_b_subgraph ← node_to_subgraph(node_b);
    if is_undefined(node_a_subgraph) and is_undefined(node_b_subgraph) then
        /* Neither node has a subgraph yet so setup brand new subgraph              */
        new_subgraph ← generate_subgraph_id();
        subgraph_to_members(new_subgraph) ← [node_a, node_b ];
        /* Update memberships                                                       */
        node_to_subgraph(node_a) ← new_subgraph;
        node_to_subgraph(node_b) ← new_subgraph;
    else if is_defined(node_a_subgraph) and is_undefined(node_b_subgraph) then
        /* A has a subgraph but B does not                                          */
        subgraph_to_members(node_a_subgraph).append(node_b);
        node_to_subgraph(node_b) ← node_a_subgraph;
    else if is_undefined(node_a_subgraph) and is_defined(node_b_subgraph) then
        /* B has a subgraph but A does not                                          */
        subgraph_to_members(node_b_subgraph).append(node_a);
        node_to_subgraph(node_a) ← node_b_subgraph;
    else if is_defined(node_a_subgraph) and is_defined(node_b_subgraph) then
        /* Both nodes have existing subgraphs                                       */
        if node_a_subgraph = node_b_subgraph then
            /* Setup new subgraph that will container merger of both node's respective
               subgraphs                                                            */
            merged_subgraph ← generate_subgraph_id();
            subgraph_to_members(merged_subgraph) ← subgraph_to_members(node_a_subgraph) +
             subgraph_to_members(node_b_subgraph);
            /* Update node subgraph memberships                                     */
            foreach member_node in subgraph_to_members(merged_subgraph) do
                node_to_subgraph(member_node) ← merged_subgraph;
            end
            /* Delete old subgraphs from map                                        */
            subgraph_to_members(node_a_subgraph) ← undefined;
            subgraph_to_members(node_b_subgraph) ← undefined;
        else
            /* Both nodes already in same subgraph so do nothing                    */
    /* Dump current state of subgraphs to histogry map                              */
    subgraphs_at_strength(edges [i]) ← subgraph_to_members;
end
```

Algorithm 1: Find neighborhood subgraphs

Figure 3.2: Find neighborhood subgraphs algorithm. By simply requiring the association strengths to be sorted, the only assumption required of the strength measure is monotonicity

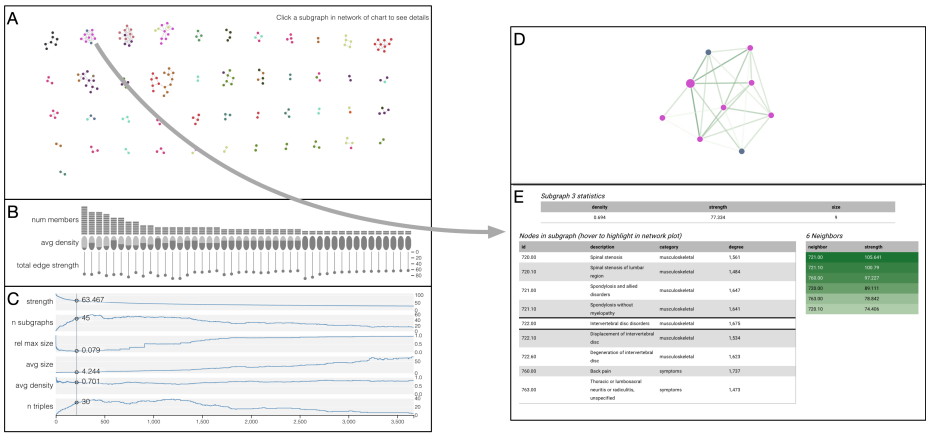

Figure 3.3: Interactive visualization of subgraph clustering results with current threshold set at the optimal threshold according to the smallest-largest rule.

3.3.2 Visualization

The subgraph-clustering algorithm results are visualized through an interactive visualization built using the javascript library D3 (Bostock et al., 2011; Luraschi and Allaire, 2020) that allows panning through and visualizing all cluster states that occur during the running of the algorithm.

At every step, all currently clustered nodes are displayed as a grid of force-directed subgraphs (De Leeuw, 1988) (figure 3.3A). Accompanying each subgraph is a set of three measures as encoded in a chart (figure 3.3B). These are the number of nodes in the subgraph, the density (number of edges at current threshold linking nodes relative to maximum possible), and the average strength of all those edges.

By adding edges in one-at-a-time, we are emulating the formation of a random graph. As the edges are added, isolated subgraphs form within the network. Unlike traditional network or heatmap visualizations, by physically separating the subgraphs, associationSubgraphs acknowledges that (at the current association strength threshold,) the nodes are conceptually isolated and should be represented as such.

Every subgraph can be selected and zoomed into by clicking, which reveals all members within the cluster (figure 3.3D), the edge strengths between them, and any further supplementary node information provided by the user (figure 3.3E).

To help explore association thresholds, a series of line-plots below the network visualization (figure 3.3C) provide summary statistics about the algorithm/cluster state at every possible cutoff. These include the number of subgraphs, the number of subgraphs with at least three members (triples), the average density of those subgraphs,

the average size of the subgraphs, and the size the largest subgraph relative to all other subgraph combined, and the current association threshold. Hovering over a state in these line plots updates the drawn network to the desired threshold. This updating is done in real-time, allowing the user to see what edges were added and how they changed the subgraph state.

3.3.3 Choosing "optimal" threshold

AssociationSubgraphs is meant to provide an exploratory view of the entire association network; this means the concept of the ideal threshold is not particularly important. However, we can draw inspiration from random-graph and percolation theory to estimate an "optimal" threshold value used as the initial point in the visualization.

When nodes are truly randomly connected, what is known as a "giant-component" forms very quickly. A giant component is an isolated subgraph containing a considerable portion (typically $n^{2/3}$ Newman (2018) chapter 15) of the network's nodes. There are three "phase-transitions" regarding the size of the largest isolated subgraph in the network relative to the number of edges (e) added.

If $e < n$, we would expect many small subgraphs with the largest having size on the order of $\log(n)$. If $e = n$, we would expect to still have many subgraphs, this time with an expected largest subgraph of size $n^{2/3}$. Last, after $e > n$, we would expect all nodes to be connected in one giant subgraph/component.

When the edges are not purely random, we expect a deviation from these patterns, and in practice, we see these deviations, with a giant-component typically forming well after the number of included edges surpasses the number of nodes.

To take advantage of this known behavior, we propose the "giant-subgraph-formation" rule for finding the optimal threshold. This rule states that an association network's optimal threshold value will be the one just before the giant component starts to form. This point is estimated by tracking the largest subgraph size relative to a threshold value of $O(n^{2/3})$. The optimal threshold can be seen as the step just before the largest subgraph exceeds this threshold.

As a giant component is defined with regards to the number of nodes in the network, with associationSubgraphs, we could use two different n's. The "local" n: n_{local} corresponds to how many nodes have currently been put into subgraphs at the given threshold (i.e., have been placed in existing subgraphs.) This means the giant-component

is relative to the currently displayed network at any given step. In contrast to the local threshold, the "global" threshold: n_{global} sets n as the total number of unique nodes present in the network. Typically we recommend using n_{local} as the difference between the two counts is frequently negligible, and, by using the local n, we remove the need to pre-calculate the number of unique nodes present in associations. Both options are available in the provided R package.

3.3.4 Simulating an association network

To simulate an association network, we define some number of variables (N_V) that are randomly assigned to some number of clusters (N_C). After every variable has a cluster assignment, all $\frac{N_V \cdot (N_V - 2)}{2}$ unique pairs of variables are given an association value drawn from a normal distribution with unit-variance and mean $\mu_0 = 0$ for pairs of variables not in the same cluster and mean $\mu_a > 0$ for pairs of variables that are in the same cluster.

3.3.5 Determining the performance of each step in the visualization

To determine how well any given state in the associationSubgraphs visualization corresponds to the true cluster structure, we compute the normalized mutual information (NMI) (Chiquet et al., 2020) between the true cluster status $\mathbf{C}$ and the current subgraph membership $\mathbf{S}_t$ for every step (3.1). NMI is the ratio of how much information is contained in both partitions of the variables. To obtain the NMI, the mutual information ($\mathbf{I}(\mathbf{C}; \mathbf{S}_t)$) between each partition is divided by the sum of the entropy present in those partitions alone ($\mathbf{H}(\mathbf{C}), \mathbf{H}(\mathbf{S_t})$). This ratio takes a value between zero (no shared information) and one (they perfectly mirror each other). The normalized aspect of this metric allows us to compare values across a different number of clusters.

$$\text{NMI}_t(\mathbf{C}; \mathbf{S}_t) = \frac{2 \cdot \mathbf{I}(\mathbf{C}; \mathbf{S}_t)}{\mathbf{H}(\mathbf{C}) + \mathbf{H}(\mathbf{S_t})} \tag{3.1}$$

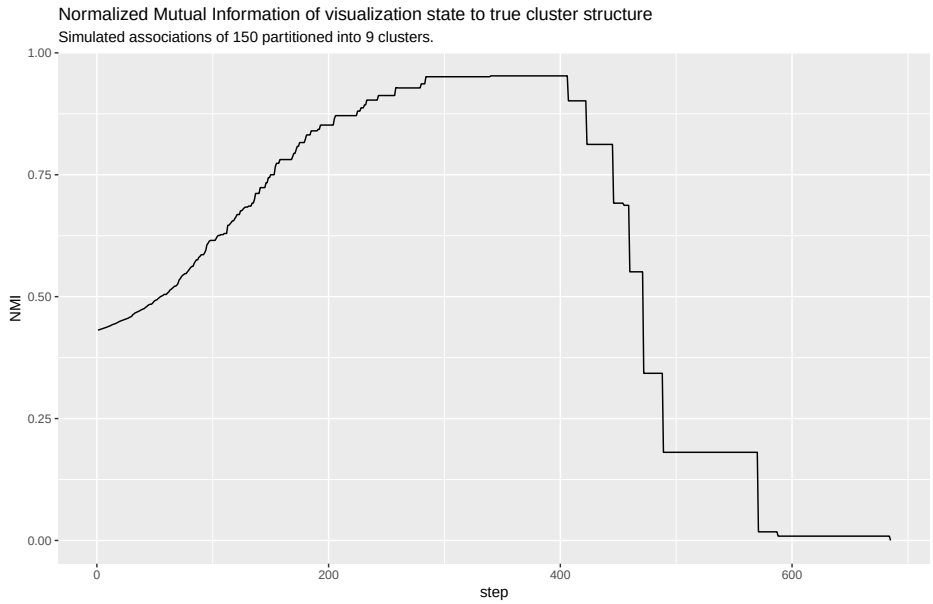

Figure 3.4: The normalized mutual information between the visualization state at a given threshold and the true cluster structure in this given simulation shows a clean peak shape with a maximum value of 0.95.

3.4 Results

3.4.1 Performance of giant-component cutoff on simulated data

The optimal cutoff suggestion is meant as just that, a suggestion. We encourage the user to explore all cutoff thresholds when using associationSubgraphs. However, the choice of the threshold is not arbitrary and, under certain assumptions, does correspond to the best cutoff position or very close to it.

To provide context for the relative associations used to access performance, figure 3.4 shows the raw performance of associationSubgraphs running on a network simulated using 3.3.4. This figure shows the associationSubgraphs algorithm reaching very close to perfect separation in the middle of the x-axis range. A behavior we see repeated in most simulations.

3.4.1.1 Visualizing the thresholds

When the network has relatively few "noise nodes," as is the case with our simulated data, then the global threshold is more appropriate as it is less likely to be triggered too early by a set of highly associated nodes occurring relatively early in the visualization.

Luckily, as Figure 3.5 shows, these two thresholds rapidly converge, so which one is

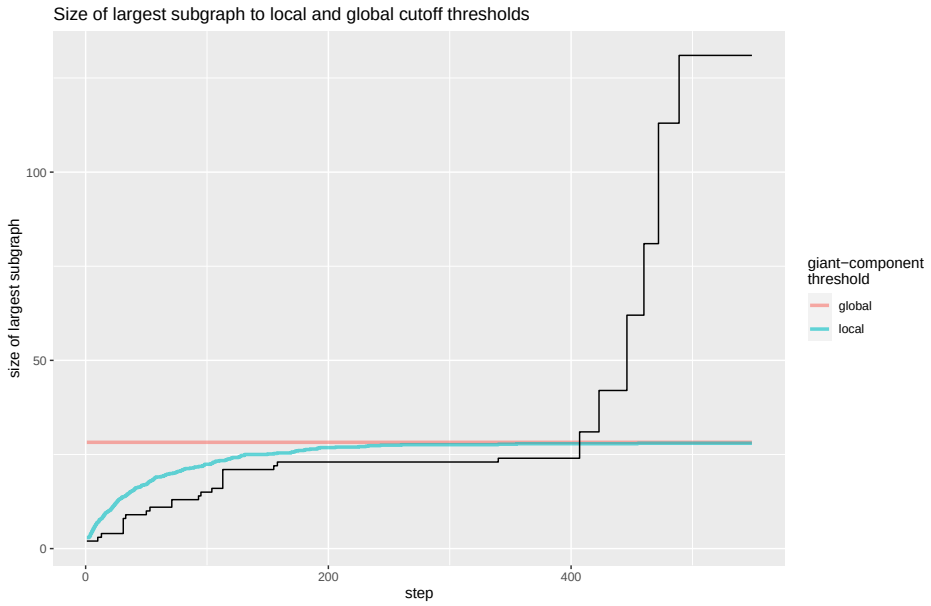

Figure 3.5: The cutoff hueristic corresponds to when the black-line (size of the largest subgraph) crosses either the local or global thresholds, indiciating a giant-component has formed. The local and global thresholds quickly converge to similar values as the number of variables seen rises; this causes the optimal thresholds according to either to be very similar.

picked is relatively unimportant. Indeed, the results for either threshold are the same step for many networks, as is the case in figure 3.6.

3.4.1.2 Scaling up simulations

The simulation described in 3.3.4 was run over a range of true cluster sizes from two to twelve, and for each size, the simulation was repeated 100 times to get a better handle on the optimal cutoff threshold's performance. The mean of the normal distribution governing the "true" associations, μ_a, was set at 3 and the number of variables fixed at 300.

As figure 3.8 shows, we see mediocre performance when the true number of clusters is small; however, the results quickly rise to excellent performance, with the median relative NMI even hitting one for both global and local cutoffs at 10 clusters.

While the assumptions made in these simulations are relatively strong (e.g., non-hierarchical structure, purely normal distributed associations, even cluster sizes), the results show that the use of giant-component formation as a threshold for an optimal cutoff has statistical merit and serves as a good starting point for the associationSubgraphs visualization. Again, the purpose of associationSubgraphs is to explore the structure dynamically, so we highly encourage the user to explore other thresholds.

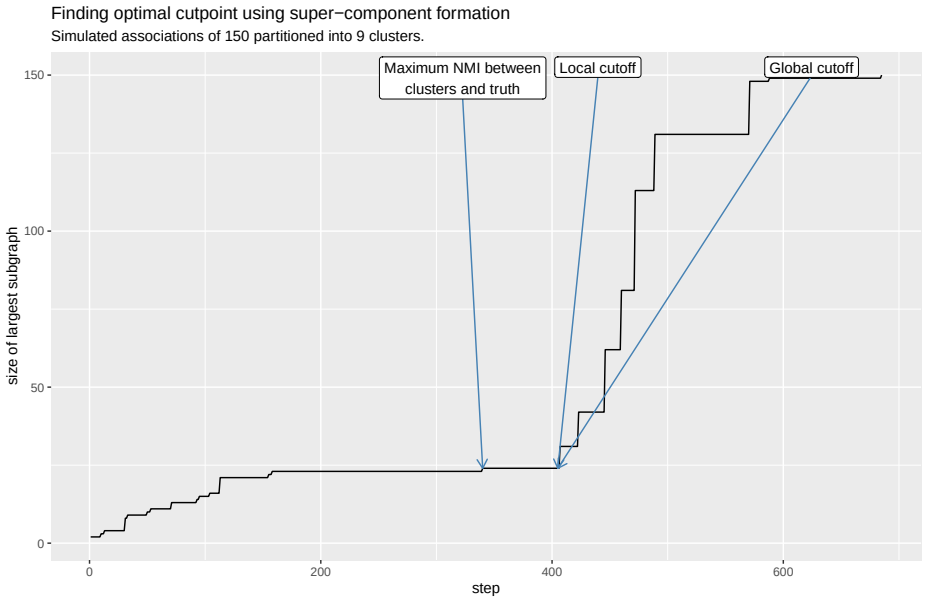

Figure 3.6: In this simulation, the local and global cutoff thresholds are very similar and fall a few steps after the true maximum NMI step as the algorithm merged two true clusters before the cutoff threshold.

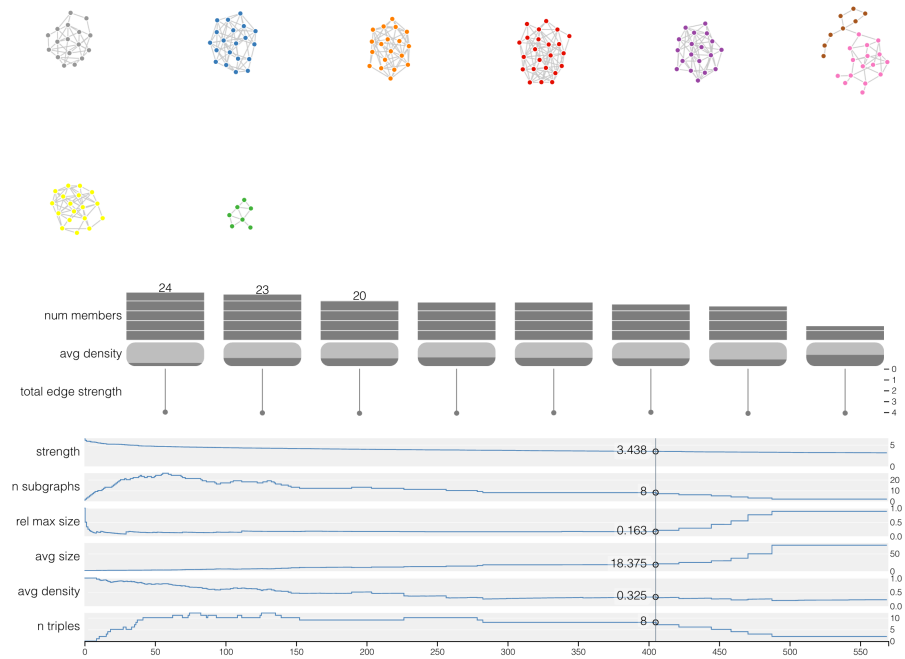

Figure 3.7: The result of running the associationSubgraphs algorithm on a single instance of the simulated association network and using the global threshold to choose the starting position. We see that the clusters are very close to perfectly resolved at the starting point.

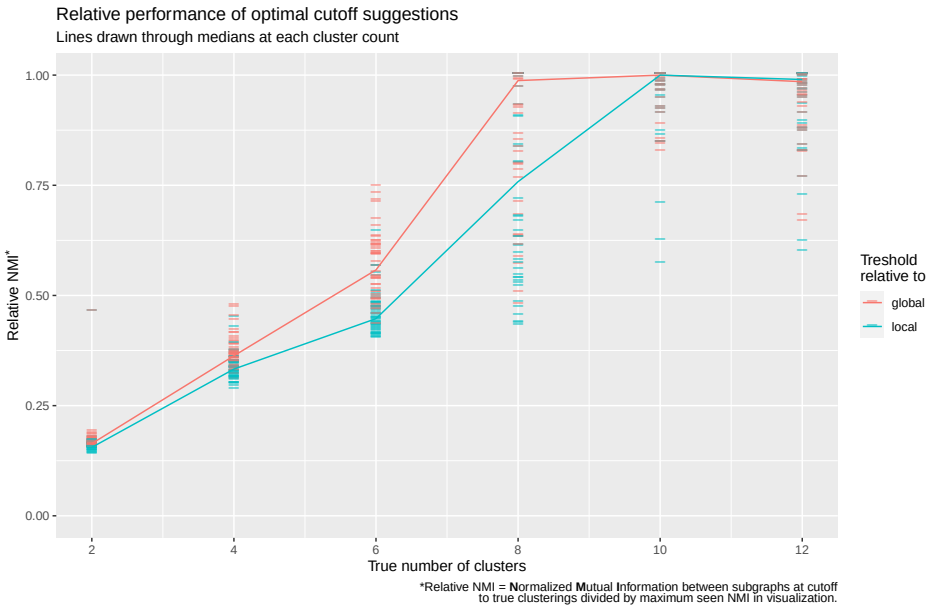

Figure 3.8: Simulations show that for non-hierarchichal cluster structure the cutoff detection heuristic works best when the number of clusters is above 6, with the relative performance in terms of NMI rising to near one for the global cutoff with the local lagging just slightly behind.

3.5 Visualizing comorbidity associations

Figure 3.3 shows the results of running the associationSubgraphs algorithm and visualization on a comorbidity network of 1,815 phenotypes as "phecodes" (Denny et al., 2010) constructed using EHR data from two large healthcare systems (chapter 4). AssociationSubgraphs provides intuitive and meaningful insights into the structure of the results. Using the described method (3.3.3) to determine an association strength cutoff returns a network with 47 isolated subgraphs. Figure 1C shows the investigation of one of the present subgraphs which includes the codes 720.00, 720.10, 721.00, 721.10, 722.00, 722.10, 722.60, 760.00, 763.00; all are back-pair related phecodes (e.g., Spinal stenosis of lumbar region: 720.10, and Back pain: 760.00.) This particular visualization and more examples are available on the associationSubgraphs R package website.

3.6 Discussion

Here, we have provided a brief introduction to the algorithm and visualization associationSubgraphs for exploring association networks. By enabling the exploration of high-dimensional association networks through interactive network-visualization guided by basic network-science theory, associationSubgraphs allows researchers to

understand their data with greater precision and intuition.

CHAPTER 4

INVESTIGATING PHENOME-WIDE COMORBIDITY LANDSCAPES ACROSS TWO LARGE-SCALE EHR SYSTEMS

4.1 Summary

Background: Large electronic health records systems have proven massively valuable for retrospective research studies, allowing the gathering of much larger cohorts than previously possible. Typically EHR-based studies are performed with a single institution's records, which, combined with the high noise levels of patient-level data, raises concerns about how robust the results are when applied outside of the original system.

Methods: Here, we build two different comorbidity networks - one for Vanderbilt University Medical Center (Vanderbilt) and one for Massachusetts General Hospital (MGH) - using covariate-adjusted logistic regression to ascertain the comorbidity strength. With network statistics, we quantify how conserved phenotype comorbidity patterns are among the two systems, and merge both networks to build a standard comorbidity network.

Results: We find high amounts of similarity between the two systems, with an overall correlation of comorbidity strengths of 0.791 (95% CI: 0.788, 0.794). Additionally, we see hints of how system specialization can affect a phenotypes' position within the overall comorbidity network. Finally, we show how these network differences and similarities can provide valuable insight for a specific phenotype (Schizophrenia.)

Conclusions: Our results show great promise for the transportability of analyses run on EHR data. The combined comorbidity network produced allows the investigation of general and robust comorbidity patterns. Further, the framework we use for the construction of this combined comorbidity network can be easily and efficiently extended as more systems' data are added.

4.2 Introduction

Electronic health records (EHR) systems provide opportunities to conduct research related to clinical phenotypes on a larger scale than ever before. Many new methods,

such as Phenome-Wide association studies (PheWAS) (Denny et al., 2010), have been developed to take advantage of these new data sources. These new methods have primarily focused on single phenotype associations, e.g., the univariate association of a given single nucleotide polymorphism (SNP) on the genome with a phenotype of interest. An area that has been relatively underserviced with EHR data is the network of phenotype comorbidities.

Research on clinical comorbidities is not a new practice; however, the previous efforts have tended to either focus on a small subset of phenotypes (Avery et al., 2011) (typically classified via chart review) or have taken the form of a meta-analysis of literature and databases (Chen and Xu, 2014; van Driel et al., 2006).

While the opportunities are far-reaching for the use of EHR data in the investigation of comorbidity patterns, it is essential to keep in mind the limitations of EHR data. EHR data, particularly the billing codes we will focus on in this paper, are highly subject to human error. A code may be left on a patient's chart from a previous visit or a different patient's records, or a provider may be biased to one code due to experience with it. Also, as ICD9 and ICD10 codes used for billing purposes, they were not designed for research, which can manifest in multiple codes encoding for the same phenotype or encompassing multiple very different phenotypes. (For a more detailed look at nuances of EHR-based research, we point the reader to (Yadav et al., 2018).) Last, billing codes' outcome space is extremely high dimensional, with roughly thirteen thousand ICD9 codes and 68 thousand of the newer ICD10 standard. This high dimensionality poses risks to statistical models both for multiple-comparisons testing and the runtime of algorithms.

Efforts have been made to address the issues inherent in ICD-code based phenotype inference. One of the most successful is the phecode (Denny et al., 2010), which was developed in conjunction with the PheWAS method. Phecodes are a hand-crafted mapping of ICD9 and ICD10 codes to 1,815 hierarchical codes constructed by MDs and Bioinformaticians. They were designed to create a more robust phenotype mapping for research purposes. In addition to the advantages of curated conceptual mappings that phecodes provide, reducing the dimension of the "phenome" space benefits statistical models. Sample sizes for a given phenotype are much larger due to combining multiple ICD9 codes to a single phecode; this means greater power in models. Importantly for comorbidity networks, the phenome's dimension reduction has a squared-effect on reducing the number of possible comorbidity patterns to be investigated. There are 2.312 billion unique comorbidity pairs of ICD10 codes but

only 1.6 million pairs of phecodes. This reduction opens up the feasibility of running models across all pairs in a reasonable amount of time.

The treatment of comorbidities among phenotypes as a network is vital to acknowledge the complexity of the phenome. Most traditional approaches to investigating a phenotype of interest involve univariate association models, either between two phenotypes or between a biomarker such as a SNP and a phenotype. This simplified view has provided valuable insights but ultimately, phenotypes do not exist in isolation. Every phenotype is influenced by - or influences - another phenotype with varying levels of intensity.

By building models that account for the phenome's network structure, we open the door to much more nuanced and powerful inferences. For instance, a biomarker strongly linked with cancer formation will often have a high association with nausea. Univariate methods simply report this association, whereas network-based methods allow the researcher to see that the comorbidity is likely driven by nausea's association with the cancer phenotype due to chemotherapy and other treatments than by a direct causal path between nausea and the biomarker.

Further, by viewing comorbidity as a network, we open the door to system-level questions. One question that can be asked is what phenotypes are the most "central" in the network. That is what phenotypes tend to be comorbid with lots of others? The network structure can also be used to find phenotypes that occupy very similar positions within the comorbidity network, potentially highlighting new phenotype targets for use in drug repurposing. Infact, the insights provided by network methods for drug repurposing have already been seen in other biomedical fields such as Pharmacology (Csermely et al., 2013).

While the potential benefits of network-driven inferences on EHR-derived phenotype comorbidity are considerable, the results are most useful if they are transferable to other systems. Due to data availability and established methods, previous work in the realm of EHR-derived phenotype models has focused principally on single healthcare systems (Hebbring et al., 2013). However, differences such as patient population demographics and primary health-insurance providers will result in differences in both what a given phenotype "means" between systems and how the comorbidity network is structured between those phenotypes. Estimating how stable these underlying networks are is critical to determining if statistical models and inferences are appropriate to be applied outside their home system.

Network analysis approaches provide valuable tools to access the differences between

these networks. These approaches allow us to make statements about how stable a given phenotype is within its network: is it comorbid with the same phenotypes, or do its relationships change? Statements can also be made about the network structure itself: are the same phenotypes "central" in each system, or do other factors such as hospital specialization or patient demographics considerably influence these patterns?

As our proposed methods only compare and merge the comorbidity networks themselves (represented as strengths of comorbidities between pairs), these networks can be generated within a system and shared without risks to patient privacy. These privacy-conscious mappings provide an opportunity to combine multiple systems' information to help smooth out the differences observed between the two systems and build more general-purpose comorbidity networks without the need for large-scale databases to aggregate individual-level data such as UKBiobank. The constructed phenome-wide general comorbidity network can be used by researchers hoping to answer questions specific to phenotypes (or groups of phenotypes) of interest within the context of their comorbidities.

4.3 Methods

4.3.1 Individual-level data

Individual-level data for 250 thousand randomly selected patients were gathered from Vanderbilt and MGH's EHR systems of 2.2 million and 1.8 million patients, respectively. The sampled patient's longitudinal records were then collapsed to the number of occurrences of ICD9 codes. These ICD9 code counts were then mapped to phecodes using the PheWAS R package (Carroll et al., 2014). Finally, each phecode was considered to have "occurred" when seen two or more times in the patient's collapsed record. Demographic data, including patient age at extraction date (regardless of mortality status), EHR age (patient age at last recorded visit), sex, race, and EHR burden (number of unique phecodes present in patients' records) were also exported for model adjustment purposes (see (4.1)).

4.3.2 GLM-based comorbidity strength

Two logistic regression generalized linear models (GLMs) were run to characterize "comorbidity strength" between two phecodes for all pairs of phecodes (A, B). One predicting the occurrence of phecode A given the occurrence of phecode B, and the

other predicting the occurrence of phecode B given phecode A. The test statistic for the conditional phecode in each model was then used to represent comorbidity strength. Both models were adjusted for patient age, EHR age, sex, and race. See (4.1) for the full specification of the model in one direction.

By using the test statistic instead effect-size, the comorbidity strength takes into account the relative sample-size of each regression, avoiding large comorbidity estimates due to random chance when very few patient's records contained both phecodes A and B. If no patients had both phecodes in their records the models were not fit.

Both effect sizes were then averaged together to create a symmetric comorbidity score for all pairs. Unless otherwise noted, this symmetric comorbidity score is what is used for the analysis in this paper.

$$\begin{aligned}\log\left[\frac{P(\text{phecode B})}{1-P(\text{phecode B})}\right] =& \alpha + \beta_1(\text{phecode A}) + \beta_2(\text{age}) + \\ & \beta_3(\text{sex}_\text{M}) + \beta_4(\text{sex}_\text{UNK}) + \\ & \beta_5(\text{race}_\text{B}) + \beta_6(\text{race}_\text{H}) + \beta_7(\text{race}_\text{I}) + \\ & \beta_8(\text{race}_\text{N}) + \beta_9(\text{race}_\text{UNK}) + \beta_{10}(\text{race}_\text{W}) + \\ & \beta_{11}(\text{EHR age}) + \beta_{12}(\log(\text{EHR burden})) + \epsilon\end{aligned} \tag{4.1}$$

Equation (4.1)*: Form of model used to infer comorbidity strength. Second model is of the same format with roles of phecode A and phecode B flipped.*

4.3.3 Comorbidity patterns:

To characterize a given phecode's comorbidity behaviors, we can export what we call the phecode's "comorbidity pattern." The comorbidity pattern is a phecode's set of all defined comorbidity strengths to other phecodes (figure 4.1). It is important to note that because some phenotypes are rarer than others, it is common for there to be no overlap between patients for two phecodes and thus no information to infer comorbidity strength; this means that rarer phecodes will have smaller comorbidity patterns.

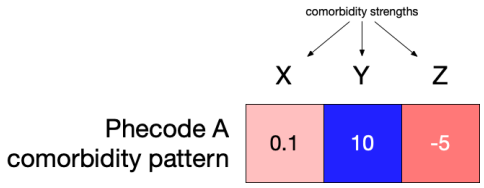

Figure 4.1: Example comorbidity pattern for a phecode to three other phecodes.

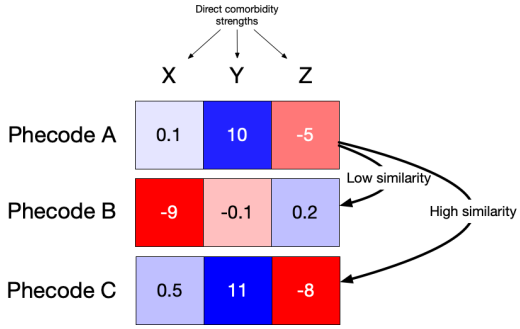

Figure 4.2: Example of three phecode comorbidity patterns where phecode A has a low similarity to phecode B but a high similarity to phecode C.

4.3.4 Comorbidity similarity

How much the comorbidity patterns of two phecodes resemble each other conveys how similar the two phecode's are in the broader comorbidity network. This measure is referred to as the "structural equivalence" (Newman, 2018, chapter 7) of the two phecodes within their network. Here, we define the similarity between the two comorbidity patterns as the Pearson correlation of their common comorbidities. Pearson correlation is used as the comorbidity strengths are - due to using the test statistic of association - gaussian distributed.

For example, as in figure 4.2, if phecode A has comorbidities of 0.1, 10, and -0.5 to phecodes X, Y, Z, respectively, and phecode B has comorbidities of 8, -0.1, and 0.2: they would have low comorbidity similarity. Whereas phecode C with comorbidity strengths of 0.5, 11, and -8 would be highly similar to phecode A.

4.3.5 Phecode comorbidity conservation

To assess how conserved a phecode's relative position is within the comorbidity network we again use a node-similarity metric. This similarity compares the comorbidity pattern of a phecode with the *same* phecode in the opposing system (see 4.3.) Again, this is calculated by taking the Pearson correlation of two vectors.

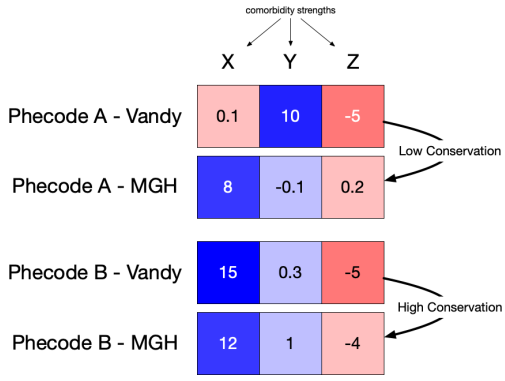

Figure 4.3: Example of two phecodes across two EHR systems where phecode A's comorbidity pattern is relatively unconserved between the systems but phecode B is highly conserved.

4.3.6 System-level comorbidity conservation

A snapshot of the conservation of the entire comorbidity network is obtained by taking a weighted average of all individual phecode comorbidity conservation values, where weights are given by the size of the shared comorbidity patterns between the two systems. By using a weighted average, rarer codes - with less certain estimates of conservation due to fewer defined comorbidity strengths - will have less impact on the system-wide conservation than more common codes. While not particularly useful in isolation, system-level comorbidity conservation provides a comparison point for future expansion of these analyses.

4.3.7 Phenotype centrality

The network statistic eigen-centrality is used to measure how "important" a phecode is within the comorbidity network. Eigen-centrality (Newman, 2018) is an "importance" measure of a node that considers the number and strength of its connections *and* who those connections are (figure 4.4). For instance, a phenotype with strong connections to rare and relatively un-comorbid phenotypes will rank lower than one with the same magnitude of connections to more commonly comorbid phenotypes such as cardiovascular disease. Phenotype-centrality for a single phenotype in isolation is rather uninformative, but when the entire network's centrality measures are investigated, centrality paints an information-rich view of the structure of comorbidities in the system.

Eigen-centrality is calculated by taking calculating the eigen-vector with the largest magnitude of the symmetric adjacency matrix (4.2).

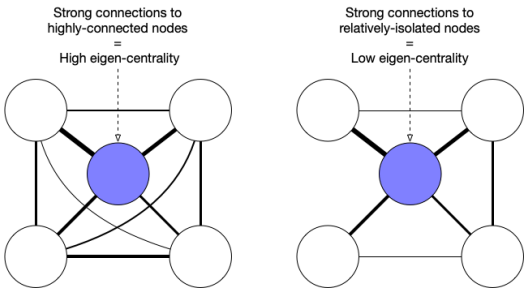

Figure 4.4: Eigen-centrality takes into account how connected - and thus central - the neighbors of a given phecode are to give a network-structure informed view into phecode centrality.

$$\text{Centrality of phecode i} = x_i \text{ s.t. } \mathbf{Ax} = \kappa\mathbf{x}; \text{ where } \kappa = \text{largest eigenvalue of } \mathbf{A} \tag{4.2}$$

4.3.8 Combined comorbidity network

Both comorbidity networks were merged at the comorbidity pair level to build a combined comorbidity network for cross-institution inferences. Every pair of phecode associations was defined as the weighted average of both system's comorbidity strength for that pair, weighted by the number of patients that shared that comorbidity with the respective network (equation (4.3)).

$$\text{Comorbidity}^{C}_{A,B} = \frac{(\text{Comorbidity}^{V}_{A,B} \cdot \#\text{ shared}^{V}_{A,B}) + (\text{Comorbidity}^{M}_{A,B} \cdot \#\text{ shared}^{M}_{A,B})}{\#\text{ shared}^{V}_{A,B} + \#\text{ shared}^{M}_{A,B}} \tag{4.3}$$

This weighted averaging gives greater emphasis on the estimates backed by a larger number of cases. Standard inverse-variance weighting is not applicable in this instance as the comorbidity strength is the test statistic and thus does not have a variance associated with it.

4.3.9 Subgraph neighborhoods

A custom hierarchical clustering algorithm and visualization were developed (4.5) to explore the patterns within the combined comorbidity network; the algorithm

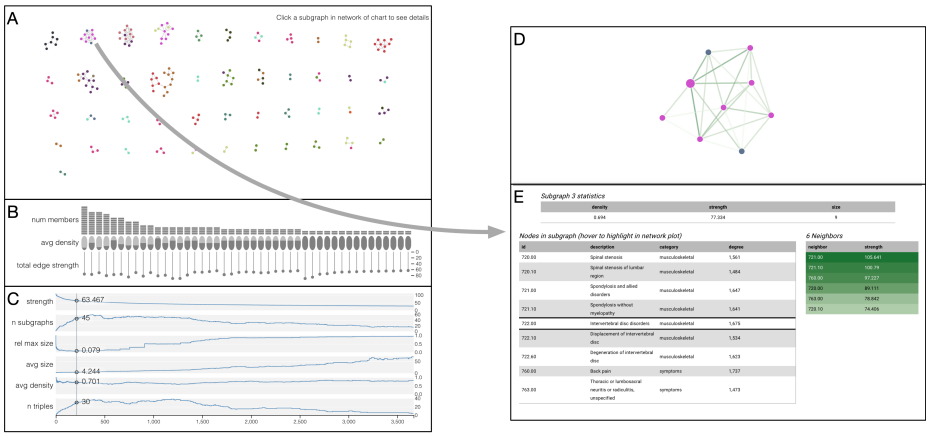

Figure 4.5: Interactive visualization of subgraph clustering results with current threshold set at the optimal threshold according to the smallest-largest rule.

is closely related to single-linkage clustering but differs philosophically by viewing nodes/phecodes that are yet to be merged with other nodes/phecodes as unclustered rather than residing within their own cluster of size one.

This difference means the results show, for a given comorbidity "threshold," what phecodes sit within neighborhoods of highly comorbid phecodes. For more details, we point the reader to chapter 3, which details the algorithm and accompanying visualization tool.

4.4 Results

4.4.1 Patient populations

4.4.1.1 Demographics

As both systems serve different populations of individuals, this means there will be differences in the population demographics of the 250k randomly sampled patients. These differences can be seen in race distributions (4.6), with the MGH population being substantially more white than the Vanderbilt population (186k vs. 154k). MGH is also more female than the Vanderbilt system (144k vs. 133k). The age distribution is also noticeably different between the systems, with Vanderbilt's population distribution being much flatter, with more young and old patients than MGH. These and other differences are expected and act as a realistic example of comorbidity transportability, as no two systems will ever have identical patient populations.

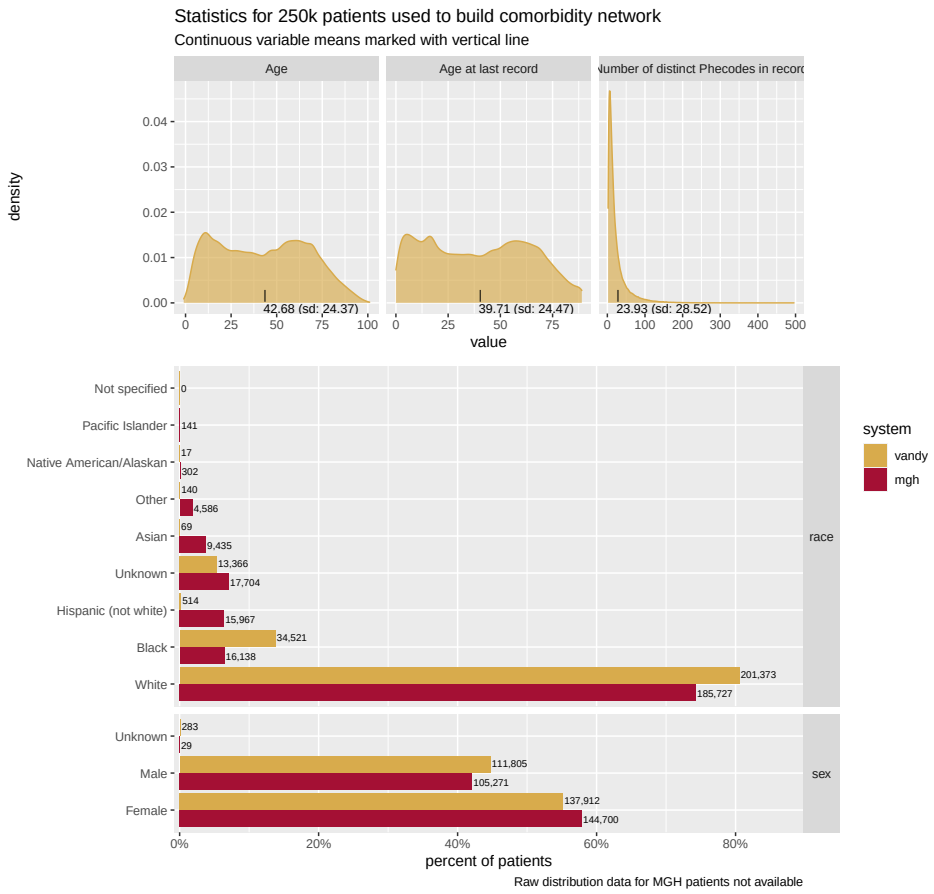

Figure 4.6: Demographic statistics for both subset of populations for both systems

Table 4.1: Age distribution differences between two systems. Note that Vanderbilt's age is regardless of mortality status.

	Counts	
Age Group	Vanderbilt	MGH
< 9	22,901	9
10 - 19	36,289	7,604
20 - 29	29,286	24,878
30 - 39	27,712	37,563
40 - 49	27,468	41,166
50 - 59	32,243	44,962
60 - 69	33,487	41,922
70 - 79	24,735	29,577
80 - 89	12,106	18,289
90 - 99	3,690	3,963
100 <	82	67
Mean age	42.7 (SD: 24.4)	52.6 (SD: 19)

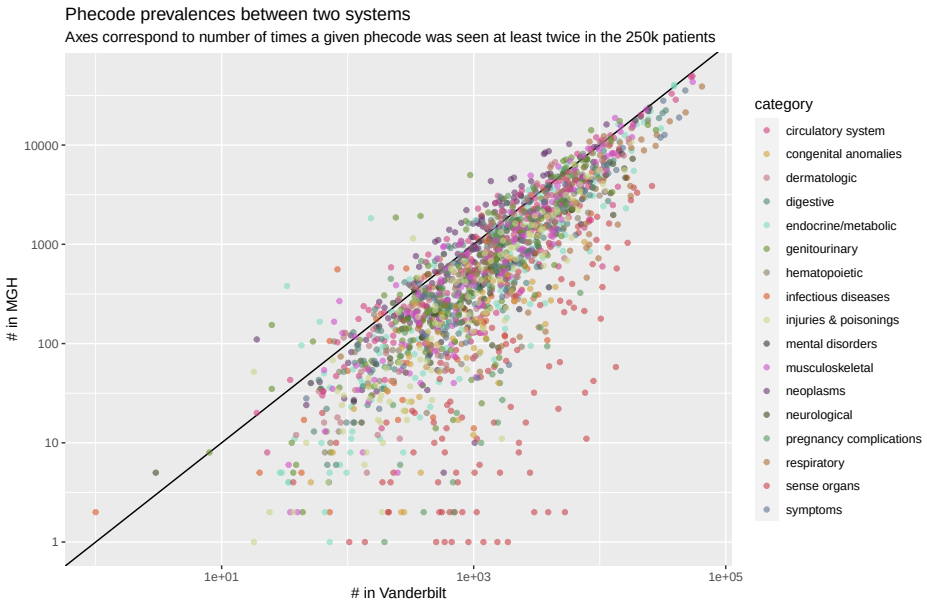

Figure 4.7: Prevalences of all phecodes across both systems show high correlation.

4.4.1.2 Phecode prevalences

Figure 4.7 compares the prevalence (or counts) of all phecodes across the systems and shows a strong agreement between the two. One instance where we see differences is the "sense organs" category, who's codes tend to have much higher levels of prevalence in Vanderbilt than MGH. This difference in counts can be attributed to the eye institute at MGH being kept separate from the main system EHR until 2008, thus limiting the accumulation of eye-related phecodes.

4.4.2 Direct comorbidity results

The results of running all possible pairs phecode's through the GLM-based comorbidity model shows a high correlation among comorbidity patterns (figure 4.8). Both populations show right-skewed distributions with very high-levels of correlation (0.791 (95% CI: 0.789 - 0.794)). This high correlation demonstrates that the comorbidity structure is preserved across the systems even when the patient populations themselves are different.

4.4.2.1 How conserved are comorbidity patterns across the two systems?

As figure 4.9 shows, conservation levels of individual phecodes (4.3.5) are varied, but only twelve have 95% confidence intervals that cover zero, or no correlation/-

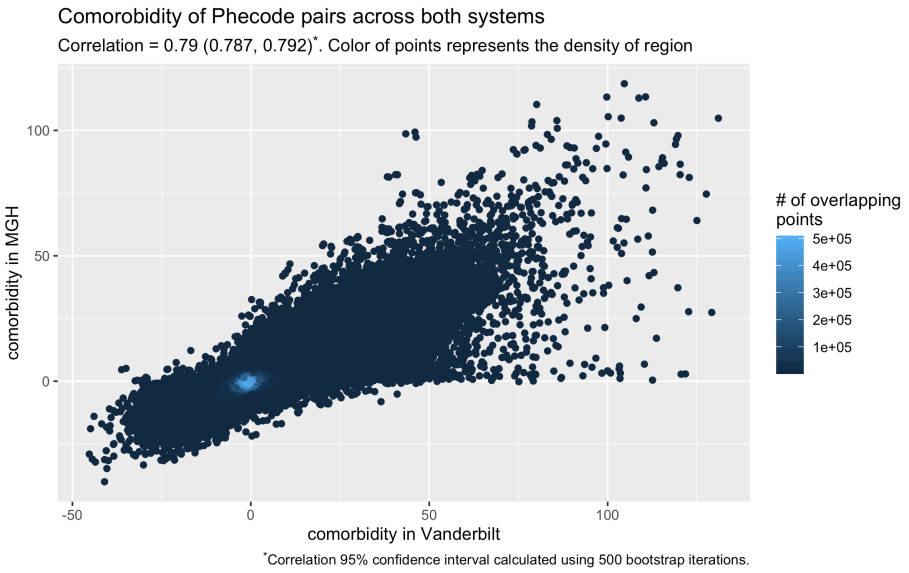

Figure 4.8: Strength of comorbidity for all common comorbidity pairs across both systems.

conservation. The median (point-estimate) conservation is 0.77, which shows high conservation of comorbidity patterns across the two systems.

Phecodes that are most conserved fall into the categories of "Circulatory system" and "musculoskeletal," as table 4.2 shows. On the other side of the spectrum, the phecodes that are the 'least' conserved - as defined by the lowest upper bound of the 95% confidence interval - are a mixture of categories. Further detail of these category differences can be seen in figure 4.11, which shows sense organ phecodes to be the least conserved on average (as would be expected with the aforementioned eye-center inclusion discrepancies in MGH) and neoplasms to be the most conserved on average.

At the individual-code level, "Primary/intrinsic cardiomyopathies" (425.10) has the highest conservation level with a lower-bound of its 95% confidence interval at 0.921. The phecode with the lowest upper bound is "Dental abrasion, erosion and attrition" (521.20) with an upper-bound of its 95% confidence interval at 0.072. The comorbidity patterns of these two extremes are shown in figure 4.10.

4.4.2.2 Phecode centrality differences

To investigate how a given phecode's comorbidity pattern is conserved across the systems and how that phecode's relative role within the broader comorbidity network is preserved, we can compare the eigen-centrality of each phecode (4.3.7) between the systems.

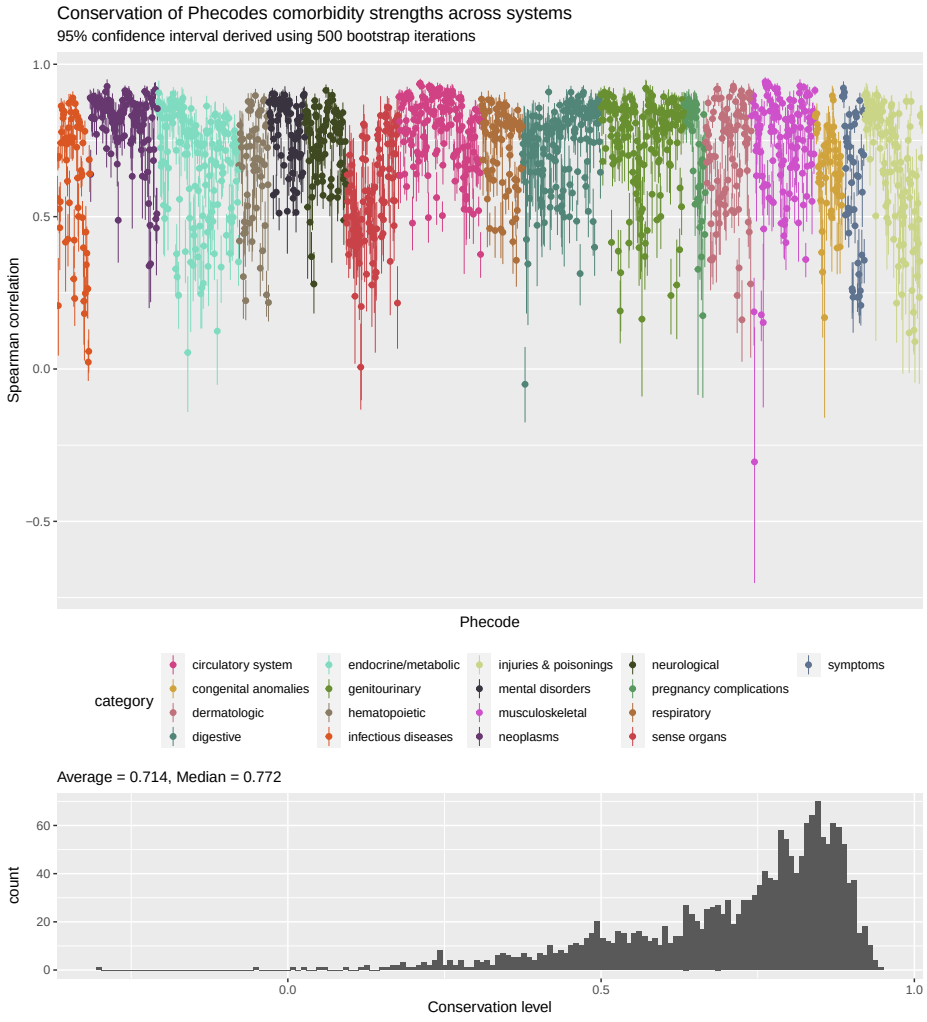

Figure 4.9: Conservation of comorbidity patterns across phenome and confidence intervals. The distribution of phecode conservation means is highly left-skewed, showing strong conservation of phecode comorbidity patterns.

Table 4.2: Most extreme conservation levels across phenome.

Phecode	category	conservation	# in common
425.10 Primary/intrinsic cardiomyopathies	circulatory system	0.938(0.921, 0.953)	1621
720.00 Spinal stenosis	musculoskeletal	0.943(0.92, 0.958)	1680
425.00 Cardiomyopathy	circulatory system	0.935(0.918, 0.949)	1634
722.00 Intervertebral disc disorders	musculoskeletal	0.939(0.917, 0.955)	1731
722.60 Degeneration of intervertebral disc	musculoskeletal	0.93(0.911, 0.945)	1699
720.10 Spinal stenosis of lumbar region	musculoskeletal	0.935(0.908, 0.952)	1629
740.90 Osteoarthrosis NOS	musculoskeletal	0.925(0.908, 0.939)	1746
427.10 Paroxysmal tachycardia, unspecified	circulatory system	0.926(0.907, 0.94)	1616
702.10 Actinic keratosis	dermatologic	0.928(0.907, 0.948)	1647
740.00 Osteoarthrosis	musculoskeletal	0.924(0.907, 0.939)	1749
521.20 Dental abrasion, erosion and attrition	digestive	-0.05(-0.175, 0.072)	246
134.00 Helminthiases	infectious diseases	0.023(-0.039, 0.096)	868
134.10 Intestinal helminthiases	infectious diseases	0.058(-0.012, 0.13)	678
710.30 Osteopathy resulting from poliomyelitis	musculoskeletal	-0.305(-0.702, 0.137)	23
367.10 Myopia	sense organs	0.006(-0.133, 0.149)	163
974.00 Poisoning by water, mineral, and uric acid metabolism drugs	injuries & poisonings	0.127(0.047, 0.214)	481
976.00 Poisoning by agents primarily affecting skin & mucous membrane, ophthalmological, otorhinolaryngological, & dental drugs	injuries & poisonings	0.09(-0.045, 0.214)	201
965.20 Antirheumatics causing adverse effects in therapeutic use	injuries & poisonings	0.119(-0.013, 0.249)	188
715.30 Spinal enthesopathy	musculoskeletal	0.178(0.091, 0.257)	277
289.90 Abnormality of red blood cells	hematopoietic	0.218(0.156, 0.276)	1271

Note:

Conservation is Pearson correlation of comorbidity vectors between Vanderbilt and MGH.

95% confidence interval (CI) calculated using 500 bootstrap iterations.

Highest lower-bound of CI determines "most conservered", lowest upper-bound determines "least conservered".

"# in common" is number of shared comorbidity strengths between systems used to calculate the correlation values.

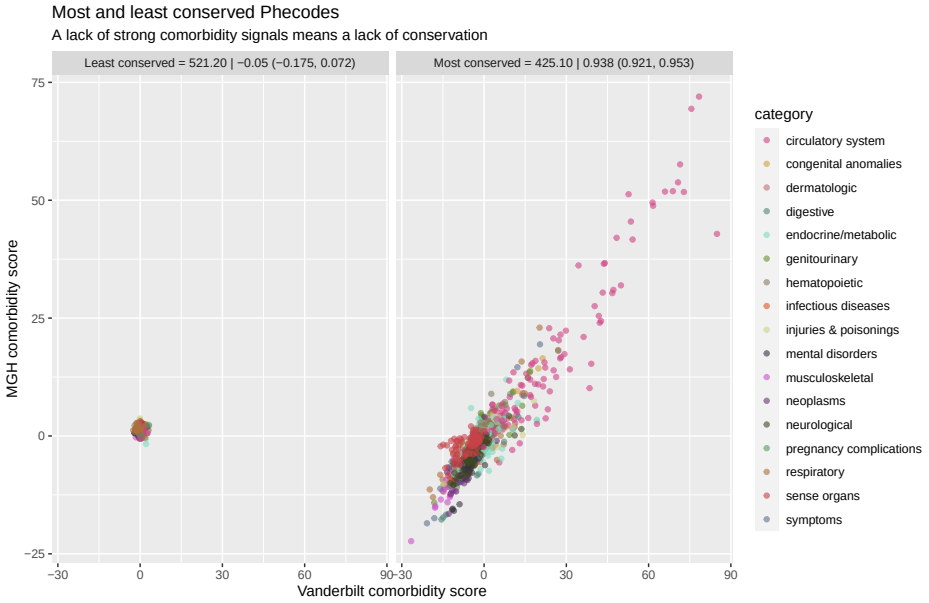

Figure 4.10: Scatterplot of comorbidity patterns of the least and most conserved phecodes show much stronger comorbidities in the more conserved code.

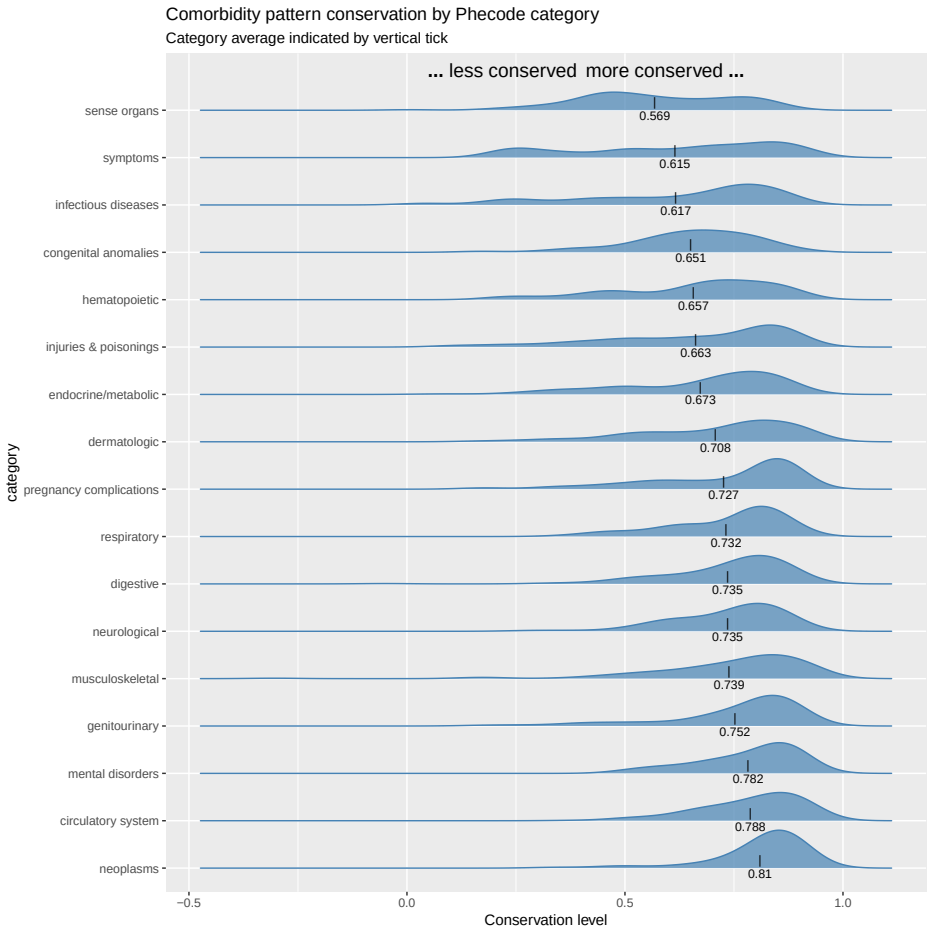

Figure 4.11: Phecode conservation patterns by category.

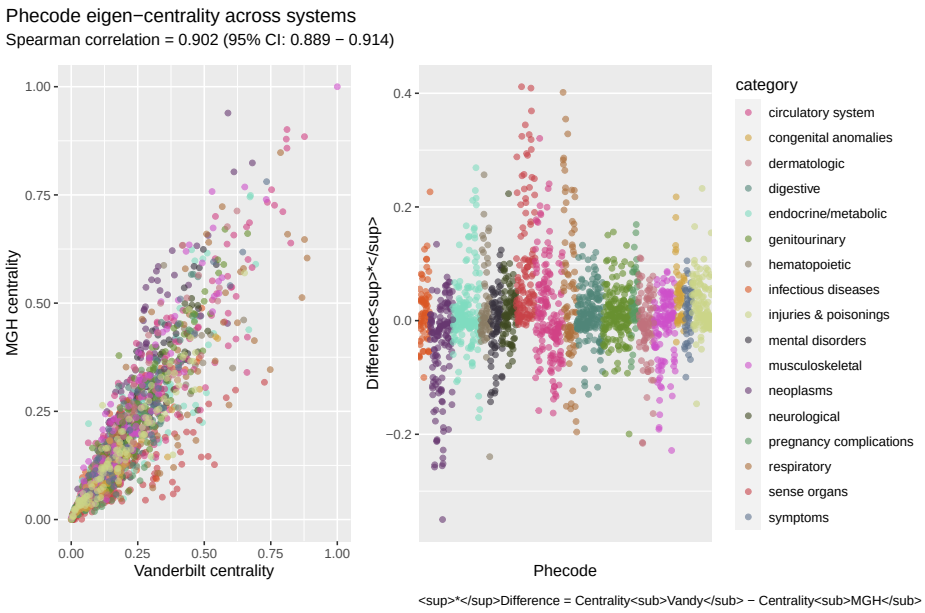

Figure 4.12: Centralities of phecodes are largely conserved across systems with some notable differences such as the sense organs category.

As 4.12 shows, there is a strong agreement between a phecode's centrality in Vanderbilt and MGH's respective comorbidity networks with a Spearman correlation of 0.902 (95% CI: 0.889 - 0.914) (Spearman used due to non-gaussian distribution of eigen-centrality).

Investigating the largest magnitude differences in centrality reveals that sense organ and respiratory codes tend to more central in Vanderbilt, whereas neoplasms make up half of the top-ten most MGH-leaning phecodes (4.3). These trends are reflected in the distribution of centrality differences by category, as shown in figure 4.13. Only three categories: neoplasms, musculoskeletal, and dermatologic are, on average, more central in MGH than Vanderbilt, with the remaining fourteen categories being, on average, more central in Vanderbilt's comorbidity network.

4.4.3 A combined comorbidity network

Comparisons of network stability across systems are valuable for understanding what comorbidity behaviors are similar or different between them, but a single network allows many exciting questions with more direct interpretability. Both systems networks were combined using (4.3) to build a more general-purpose "combined comorbidity network." As seen in figure 4.14, the resulting distribution of all comorbidities

Table 4.3: Sense organs and respiratory phecodes are more central in Vanderbilt's comorbidity network whereas neoplasms, musculoskeletal and dermatologic are more central in MGH. Centrality is normalized eigen-centrality of node given symmetric comorbidity associations. Most central node has eigen-centrality of 1.

		centrality		
Phecode	category	Vanderbilt	MGH	difference
366.00 Cataract	sense organs	0.538	0.127	0.412
381.00 Otitis media and Eustachian tube disorders	sense organs	0.726	0.317	0.409
464.00 Acute sinusitis	respiratory	0.748	0.346	0.402
381.20 Eustachian tube disorders	sense organs	0.515	0.146	0.369
476.00 Allergic rhinitis	respiratory	0.867	0.513	0.355
371.00 Inflammation of the eye	sense organs	0.489	0.144	0.344
483.00 Acute bronchitis and bronchiolitis	respiratory	0.593	0.265	0.329
379.00 Other disorders of eye	sense organs	0.398	0.071	0.327
366.20 Senile cataract	sense organs	0.383	0.059	0.324
414.00 Other forms of chronic heart disease	circulatory system	0.661	0.340	0.321
694.20 Other dyschromia	dermatologic	0.417	0.633	-0.216
198.40 Secondary malignant neoplasm of liver	neoplasms	0.324	0.552	-0.228
740.90 Osteoarthrosis NOS	musculoskeletal	0.529	0.758	-0.228
289.40 Lymphadenitis	hematopoietic	0.348	0.587	-0.239
198.20 Secondary malignancy of respiratory organs	neoplasms	0.380	0.621	-0.242
198.30 Secondary malignant neoplasm of digestive systems	neoplasms	0.302	0.554	-0.252
165.10 Cancer of bronchus; lung	neoplasms	0.320	0.574	-0.254
196.00 Radiotherapy	neoplasms	0.281	0.536	-0.255
165.00 Cancer within the respiratory system	neoplasms	0.327	0.585	-0.258
198.00 Secondary malignant neoplasm	neoplasms	0.589	0.939	-0.350

Note:
difference = Vanderbilt centrality - MGH centrality

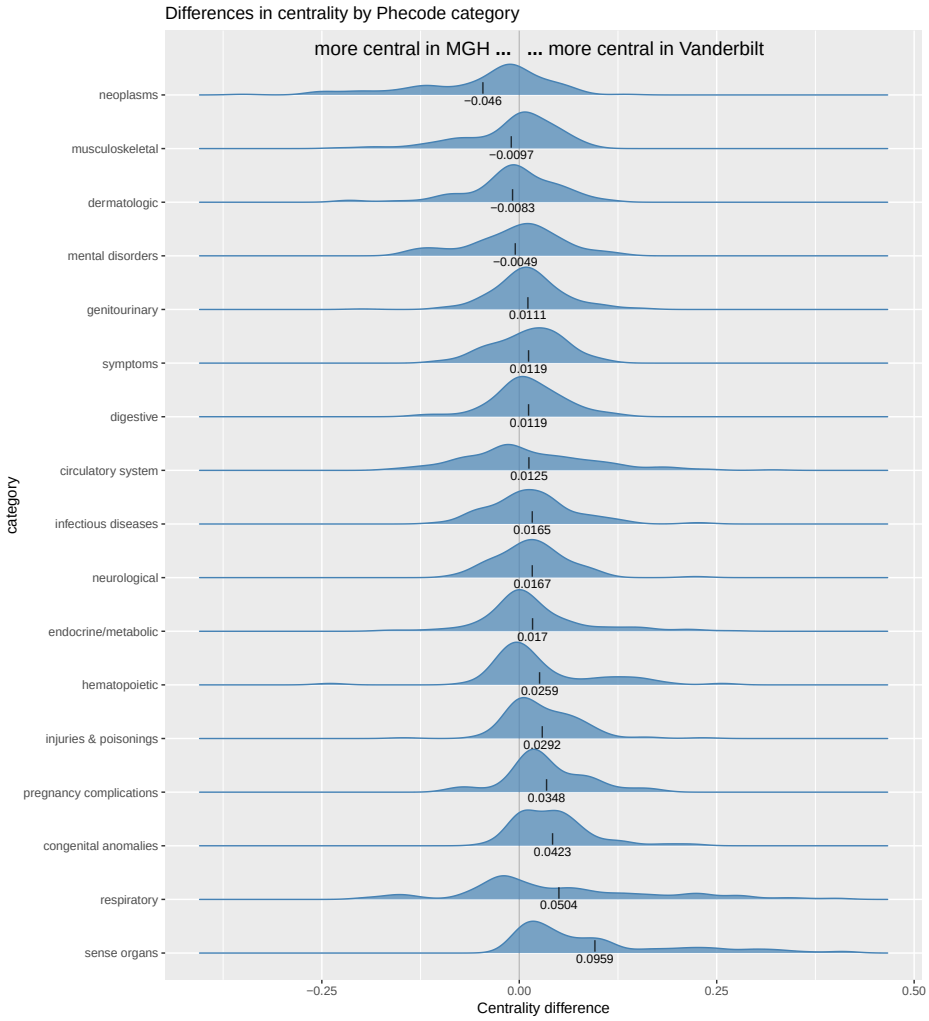

Figure 4.13: Distributions of centrality differences by category.

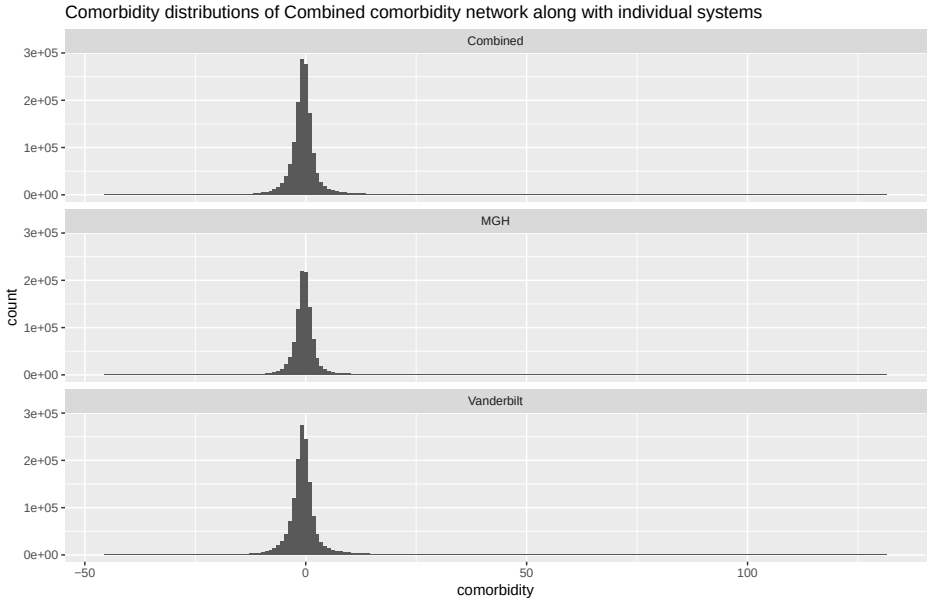

Figure 4.14: Combined comorbidity distribution is a weighted average of either system and, as a product of two gaussian distributions, maintains a gaussian distribution.

retains the same shape as each system individually: being centered around zero with a long tail of positive-comorbidity values.

4.4.3.1 UMAP projections

One way of viewing the entire comorbidity network as a snapshot is by using dimensionality reduction using the absolute comorbidity value as the "distance." Figure 4.15 shows the result of running the dimensionality reduction technique Uniform Manifold Approximation Protocol (UMAP) (Mcinnes and Healy, 2018) on the combined comorbidity network's direct comorbidity values and the phecode similarity values (4.3.4). Both networks show a strong preservation of category structure.

4.4.3.2 Neighborhood clustering

Using the association network visualization algorithm associationSubgraphs from chapter 3 provides further insights into the structure of the results. Using the recommended cutoff from (Strayer et al., 2020) to determine a comorbidity strength cutoff returns a cutoff value of 61.9454517, at which point there are 53 isolated subgraphs. One of the present subgraphs includes the codes *720.00, 720.10, 721.00, 721.10, 722.00, 722.10, 722.60, 760.00, 763.00* which are all back-related phecodes (e.g. Spinal stenosis of lumbar region: 720.10, and Back pain: 760.00.) For further

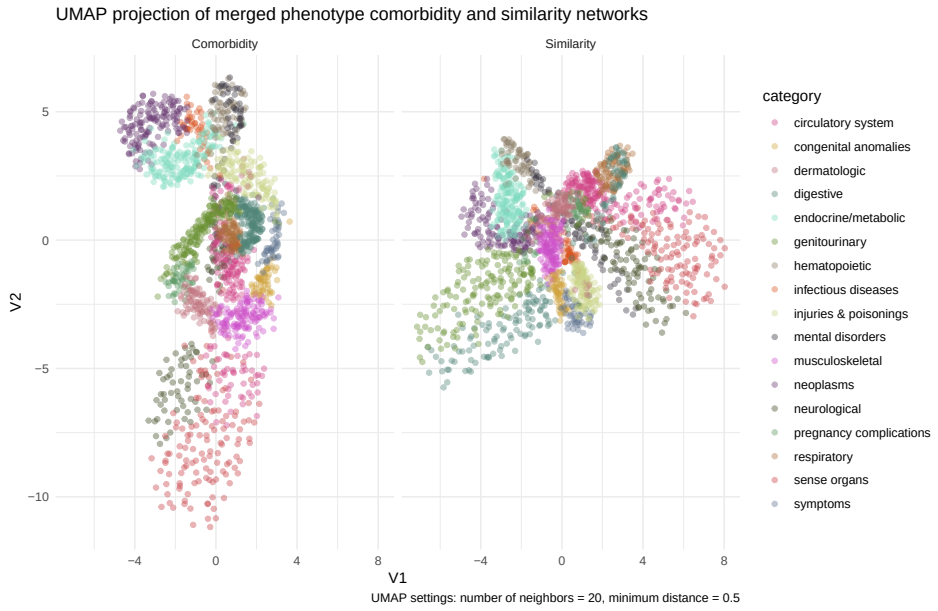

Figure 4.15: Dimension reduction of combined comorbidity network via direct comorbidity and similarity via UMAP provides clear separation of phecodes into their respective categories.

investigation of these clustering results, see figure 4.5 and the application outlined in 4.4.4.

4.4.4 App to explore phecode comorbidity

We have provided a web-application to explore a preferred phecode's comorbidity pattern and position within the broader comorbidity network. The application is available at prod.tbilab.org/comorbidity_network_explorer. This application allows the user to input a phecode of interest and returns a series of plots and tables that explore phecode's position within the combined comorbidity networks. The following section details the results provided with this application.

4.4.5 Investigating 295.1 - Schizophrenia

Here we look in more detail at phecode 295.1, or "Schizophrenia" to demonstrate comorbidity networks' utility for a specific phecode of interest.

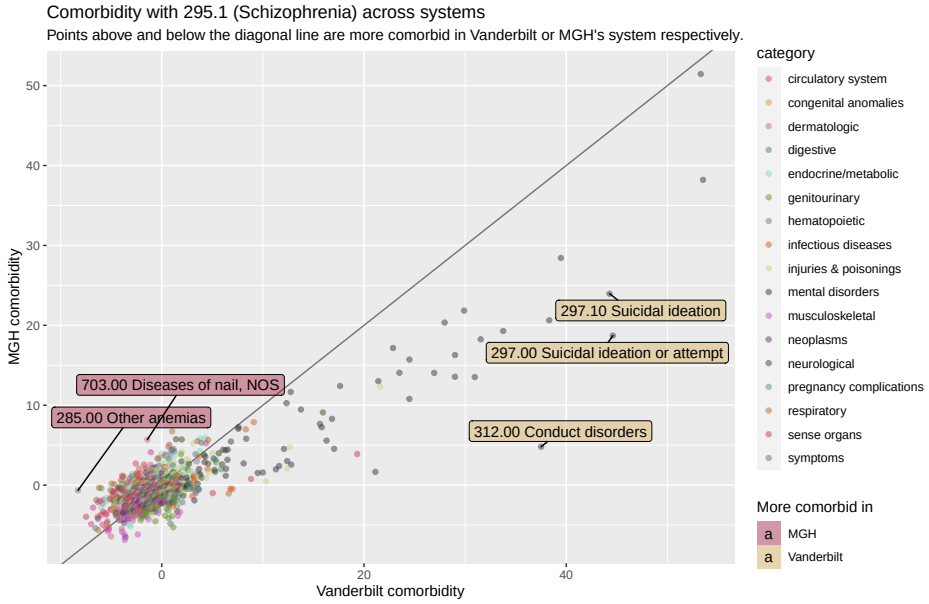

Figure 4.16: Comorbidity patterns for both systems with 295.1 are largely inline with eachother with a phecodes in the more-extreme end of comorbidity (such as 312: Conduct disorders) slighlty more comorbid for Vanderbilt than MGH.

4.4.5.1 Differences between systems

Before we look at the position of 295.1 in the combined comorbidity network, we first investigate how its comorbidity patterns match or differ for it in Vanderbilt and MGH's separate networks. Figure 4.16 shows that there is a large agreement between the two systems (conservation values: 0.852 (95% CI: 0.792 - 0.895). The top differences, as seen in table 4.4, show some phecodes that have negative comorbidity scores in MGH, such as 458 (hypotension), are positively comorbid in Vanderbilt. However, none of these differences sit in the tail of either comorbidity distribution; this contrasts with the Vanderbilt enriched codes, which tend to be highly comorbid in both systems.

4.4.5.2 Combined comorbidity network

The combined comorbidity network shows us an expected enrichment of mental disorders codes in 295.1's comorbidity pattern (figure 4.17 and table 4.5). Outside of mental disorders, 496.20 (Chronic bronchitis) stands out as an interesting comorbidity, potentially linked to high smoking prevalence among patients with Schizophrenia (de Leon et al., 1995; Lohr and Flynn, 1992; Hughes et al., 1986).

Table 4.4: The top differences in comorbidity show mental disorders codes tend to be more comorbid in Vanderbilt.

		comorbidity		
phecode	category	Vanderbilt	MGH	difference
More comorbid in MGH				
285.00 Other anemias	hematopoietic	-8.29	-0.66	-7.62
703.00 Diseases of nail, NOS	dermatologic	-1.43	5.69	-7.12
513.00 Respiratory abnormalities	respiratory	-6.21	-0.33	-5.88
496.00 Chronic airway obstruction	respiratory	1.02	6.73	-5.71
585.00 Renal failure	genitourinary	-5.99	-0.39	-5.60
458.00 Hypotension	circulatory system	-2.58	2.71	-5.29
458.10 Orthostatic hypotension	circulatory system	-1.08	4.11	-5.20
479.00 Other upper respiratory disease	respiratory	-4.96	0.19	-5.15
More comorbid in Vanderbilt				
368.91 Psychophysical visual disturbances	sense organs	19.32	3.89	15.43
300.00 Anxiety disorders	mental disorders	28.99	13.56	15.44
300.90 Posttraumatic stress disorder	mental disorders	30.96	13.52	17.44
292.60 Hallucinations	mental disorders	38.32	20.63	17.69
291.10 Transient mental disorders due to conditions classified elsewhere	mental disorders	21.12	1.66	19.46
297.10 Suicidal ideation	mental disorders	44.28	23.96	20.32
297.00 Suicidal ideation or attempt	mental disorders	44.60	18.71	25.89
312.00 Conduct disorders	mental disorders	37.51	4.81	32.71

Note:
Comorbidity defined as averaged Z-score of comorbidity regression.

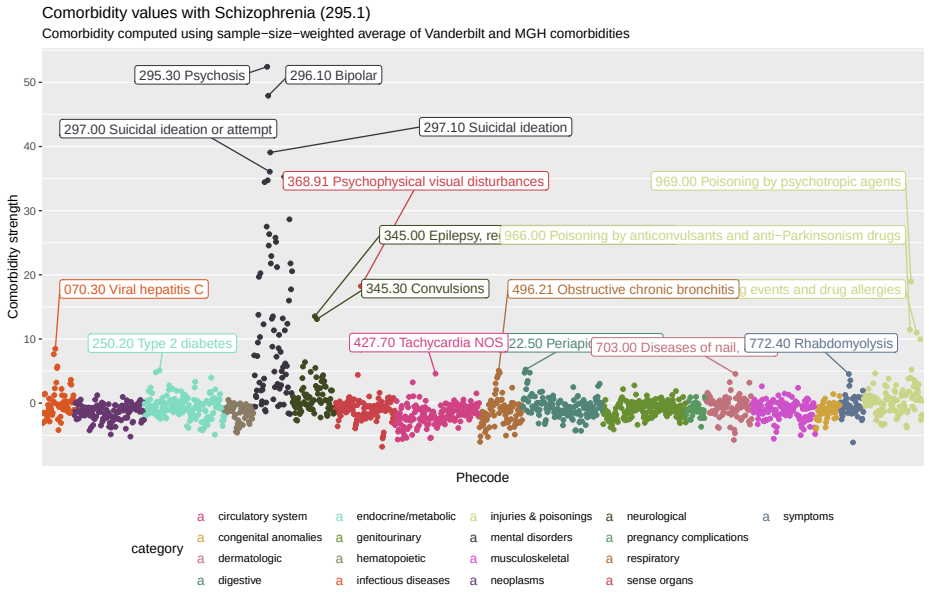

Figure 4.17: The combined comorbidity network shows high comorbidity with many expected phecodes such as 295.1: Bipolar; but also with some surprising results such as 772.4: Rhabdomyolysis.

Table 4.5: The top most comorbid phecodes with 295.1 in the combined comorbidity network are all in the mental disorders category.

phecode	category	comorbidity	similarity
295.30 Psychosis	mental disorders	52.43	0.83
296.10 Bipolar	mental disorders	47.92	0.84
297.10 Suicidal ideation	mental disorders	39.07	0.85
297.00 Suicidal ideation or attempt	mental disorders	36.08	0.79
312.00 Conduct disorders	mental disorders	35.29	0.81
296.00 Mood disorders	mental disorders	34.73	0.69
292.60 Hallucinations	mental disorders	34.44	0.83
316.00 Substance addiction and disorders	mental disorders	28.65	0.74

4.4.5.3 Contrasting comorbidity with Polygenic Risk Score

Comorbidity patterns on their own are interesting indicators of clinical-level behavior; however, one area further area of interest is how these high-level outcomes correlate with genetic signals. Figure 4.18 shows that phecodes statistically significantly associated with a Schizophrenia-pinned PRS from (Zheutlin et al., 2019). The figure shows that phecodes significantly associated with the PRS score are more correlated with their respective comorbidity values than those that were not statistically significant (although these differences themselves were *not* statistically significant). As can be seen in table 4.6, there are some strong agreements, such as 296.10 (Bipolar), which has a combined comorbidity value of 47.916 and a PRS log-odds of 0.191 (95% CI: 0.148 - 0.235).

4.5 Discussion

As EHR's become ever more prevalent, opportunities for never-before-possible research questions open up. The type of data available and the questions asked differ from traditional clinical and translational data science and statistics and require methods that fit the data and answer the right questions. By building comorbidity networks out of large amounts of patient-level data using statistically rigorous methods, we have provided a unique viewpoint into how phenotypes interact and co-occur with each other. The use of phecodes to reduce the dimensionality of the comparison space keeps the results both interpretable and more research-focused than the use of raw ICD9 or ICD10 codes. Further, by comparing two independent large-scale EHRs from both geographically and demographically distinct hospitals, we show that these

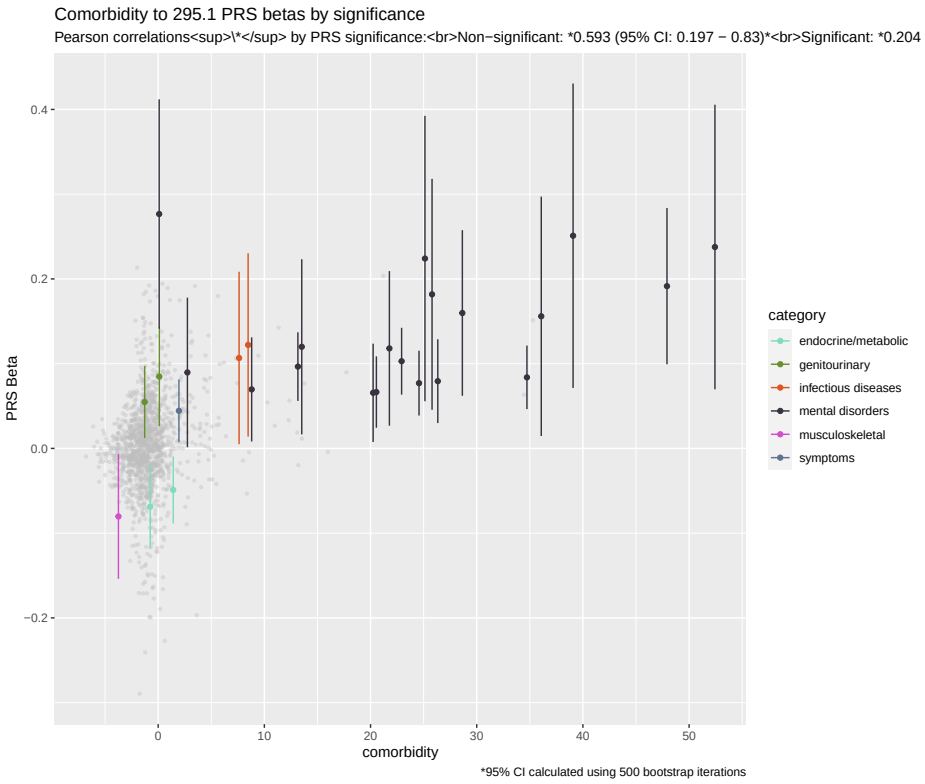

Figure 4.18: Phecodes that are statistically significantly associated with a Schizophrenia PRS score have a greater correlation with comorbidity than the rest of the phecodes within the system.

Table 4.6: Strong associations with the Schizophrenia PRS tend to imply strong comorbidity values. A notable example is 727.1 which is negatively associated with the PRS score and has an averaged comorbidity Z of -3.74.

phecode	category	PRS beta (95% CI)	comorbidity
727.10 Synovitis and tenosynovitis	musculoskeletal	-0.08 (-0.154, -0.007)	-3.741
599.00 Other symptoms/disorders or the urinary system	genitourinary	0.055 (0.012, 0.097)	-1.270
278.11 Morbid obesity	endocrine/metabolic	-0.069 (-0.118, -0.019)	-0.746
295.00 Schizophrenia and other psychotic disorders	mental disorders	0.277 (0.141, 0.412)	0.095
599.30 Dysuria	genitourinary	0.085 (0.026, 0.143)	0.109
278.10 Obesity	endocrine/metabolic	-0.049 (-0.089, -0.009)	1.418
798.00 Malaise and fatigue	symptoms	0.044 (0.007, 0.081)	1.953
292.30 Memory loss	mental disorders	0.09 (0.001, 0.178)	2.754
070.00 Viral hepatitis	infectious diseases	0.107 (0.005, 0.209)	7.620
070.30 Viral hepatitis C	infectious diseases	0.122 (0.014, 0.23)	8.471
300.11 Generalized anxiety disorder	mental disorders	0.07 (0.008, 0.131)	8.805
300.10 Anxiety disorder	mental disorders	0.097 (0.056, 0.137)	13.159
300.12 Agorophobia, social phobia, and panic disorder	mental disorders	0.12 (0.017, 0.223)	13.522
292.00 Neurological disorders	mental disorders	0.066 (0.008, 0.124)	20.243
318.00 Tobacco use disorder	mental disorders	0.067 (0.024, 0.109)	20.550
317.00 Alcohol-related disorders	mental disorders	0.118 (0.027, 0.209)	21.777
300.00 Anxiety disorders	mental disorders	0.103 (0.063, 0.142)	22.927
296.20 Depression	mental disorders	0.077 (0.039, 0.115)	24.566
301.00 Personality disorders	mental disorders	0.224 (0.056, 0.392)	25.125
300.90 Posttraumatic stress disorder	mental disorders	0.182 (0.046, 0.318)	25.792
296.22 Major depressive disorder	mental disorders	0.079 (0.03, 0.129)	26.334
316.00 Substance addiction and disorders	mental disorders	0.16 (0.062, 0.258)	28.647
296.00 Mood disorders	mental disorders	0.084 (0.046, 0.121)	34.729
297.00 Suicidal ideation or attempt	mental disorders	0.156 (0.015, 0.297)	36.078
297.10 Suicidal ideation	mental disorders	0.251 (0.071, 0.431)	39.074
296.10 Bipolar	mental disorders	0.191 (0.099, 0.284)	47.916
295.30 Psychosis	mental disorders	0.238 (0.07, 0.406)	52.433

Note:

Sorted by from smallest to largest comorbidity values.

PRS beta is log-odds of phecode occuring given increase of one unit of PRS. Confidence interval is Bonferroni-adjusted.

comorbidity networks provide meaningful and transferable insights.

There are limitations to our analysis unavoidable due to the problem space. First, due to the need to preserve patient privacy, the direct sharing of patient-level data is not possible or advisable; thus, the combination of comorbidity networks must be done as a meta-analysis.

A full investigation of what patterns of comorbidity are due to the structure of ICD codes along with their phecode mappings is currently unfeasible due to computation times. The computation of each system's network took one week on a cluster with 100 CPU cores. The simulations needed to fully characterize the structural role of coding fully would take too long with current (2020) computer architectures. However, we point the reader to the 6 where we use a null-model simulation to show that the impact of the hierarchical structure of phecodes on the conservation of the comorbidity networks, while not zero, is minimal.

The use of the test-statistic, or "Z-score," as a comorbidity strength measure, is desirable because it encapsulates the directionality and the uncertainty of comorbidity without requiring two measures as a pure effect-size would. However, it does throw away information on the exact effect-size, along with being a one-mode projection of a truly bipartite network of patients and phenotypes.

Finally, the concept of a true "comorbidity network" is one without ground truth. Unlike many statistical models, where a real underlying association or metric is being estimated, a network structure is a product of the question asked. The interplay between question and network allows great freedom to pose important questions but limits the ability to test the assumptions empirically, relying on the analyst to ask well-formed and scientifically meaningful questions. We point the reader to chapter one of the textbook Networks, by Mark Newman (Newman, 2018) for further reading on this subject.

While the network comparison results show great promise regarding the stability within individual networks, the merger of the two system's networks to a combined-comorbidity network allows the inference of general location-independent comorbidity patterns. The inferences drawn from this network via the included web-application will allow researchers to easily explore patterns of their phenotype of interest, providing a phenome-level view not previously accessible. The framework used to combine both systems is easily extendable to further systems, allowing the expansion of this work to develop evermore robust and valuable comorbidity networks.

CHAPTER 5

CONCLUSION

In Chapter 2, we showed that by merging the established PheWAS method with interactive visualizations of the underlying patient-phenotype network, researchers have greater power to explore and understand their results. While the PheWAS-ME application is available to use online without any coding necessary, we have also provided an R package with an easy-to-use modular framework self-hosting sensitive data and expanding PheWAS-ME applications.

Chapter 3 drew upon network analysis methods and graph algorithms to create a new method for exploring high-dimensional association networks such as the phenotype comorbidities. Previous methods to visualize association networks are either meant for much lower-dimensional spaces or only provide structure summaries such as dendrograms, making them unsuitable for deeply exploring patterns. Our approach, based on isolated subgraphs and interactive association thresholds, keeps the representation of the results transparent and straightforward. By pairing this simplicity with fast graph-algorithms, associationSubgraphs empowers analysts to rapidly and thoroughly explore their results. The work is accompanied by an R package with functions for creating and embedding associationSubgraph visualizations into reports and websites to make its use as easy as possible.

Chapter 4 used network statistics and visualizations to quantify comorbidity network stability across two independent EHRs. Besides providing a first-of-its-kind analysis of comorbidity structure between multiple systems, we derived new and interpretable network methods for characterizing those comorbidity networks' topology. After demonstrating substantial conservation in comorbidity structure across the two systems, we constructed a merged network and used it to explore the comorbidity patterns of Schizophrenia: providing a framework that can be repeated using an interactive web application provided. This work serves as a boost of confidence for the validity of billing-code-derived phenotypes for research. It also provides a framework for further expanding the combined general-purpose comorbidity network with data-privacy-respecting individual networks, meaning sensitive patient-level data need never leave the home institution.

Altogether, this dissertation provides a glimpse at the utility provided by merging traditional biostatistics with network analysis methods, algorithms, and visualizations. This combination is critical to exploring and understanding the correlated outcome space of clinical phenotypes. We hope that this work provides both foundation and inspiration to build new methods that further characterize and consider comorbidities when performing analyses.

CHAPTER 6

APPENDIX: SIMULATING COMORBIDITY ASSOCIATIONS

A natural question arising from such high correlations between associations in the two systems seen in chapter 4 is how much of that correlation is due entirely to how phecodes (and by extension, billing codes) are structured? One vital benefit of phecodes is their hierarchical structure. Starting at the round integer level and descending in specificity in the tens and then thousands place. This structure means that one would expect a high association between two phecodes that sit in-line on the hierarchical structure, e.g., 345.00 and 345.10 should be more associated than 345.00 and 228.12.

Here we simulate random patient data from 250,000 individuals to assess how much the hierarchical structure inherent in the phecode definitions drives the correlation between independent systems.

6.1 Simulation procedure

6.1.1 Generating patient data

The following steps were taken to simulate a patient's "phenome" with the hierarchical structure of phecode's present. First, the patient's base phenome is drawn with independent Bernoulli trials with probability of success P_c for all possible phecodes. The hierarchical structure is then added to the patient's phenome by "rolling up" all the phecodes present in their base phenome. This rollup is done by adding all phecodes above a given phecode in the hierarchical structure to the patient's phenome. For example, if a patient has phecode 295.12, both 295.1 and 295.00 are added to their phenome vector. This procedure is then repeated until the desired population size of N_p is achieved.

The probability of any phecode occurring in a patient's phenome: P_c was set at **0.01**. This value is derived from the average number of unique phecodes seen in patients in the observed systems dataset to keep the simulations as accurate to the real data as possible. It is important to note that, as the rollup adds more phecodes to a patient's phenome, this is a slight over-estimate.

Table 6.1: Statistics for comorbidity z values for simulated association pairs

system	mean	sd	min	median	max
Simulated	0.005	1.009	-3.301	-0.016	4.051
MGH	3.454	9.720	-12.220	0.266	92.816
Vanderbilt	5.142	10.389	-9.162	1.287	127.653

For computational reasons, the possible phecodes were limited to the "neoplasms" section. Neoplasms were chosen as they showed high conservation between the two systems, meaning that if the hierarchical structure is responsible, they will be an excellent candidate to reflect this.

6.1.2 Calculating the association network

The model fit to infer comorbidity of phecode pairs in simulations ((6.1)) follows the same form as (4.1) except for covariate adjustment. Since the occurrence of any phecode in a patient's phenome is uniform, we drop the covariate adjustments present in the original models we effectively "know" that the effect sizes are zero.

As in original model, the association of A given B uses the same format with roles of phecode A and phecode B flipped. These two models are then averaged together to produce a symmetric comorbidity strength.

$$\log\left[\frac{P(\text{phecode B})}{1-P(\text{phecode B})}\right] = \alpha + \beta_1(\text{phecode A}) + \epsilon \tag{6.1}$$

6.2 Results

6.2.1 Association distributions

The distribution of all associations for simulated pairs is, as expected, normally distributed. As table 6.1 shows, there is a slight positive bias (mean = 0.005) as expected from the structure, but this positive leaning is much smaller that either system (Vanderbilt mean = 5.142, MGH mean = 3.454). Further, figure 6.1 shows that both real systems exhibit longer tails than the much more clean bell-shape of the simulated data.

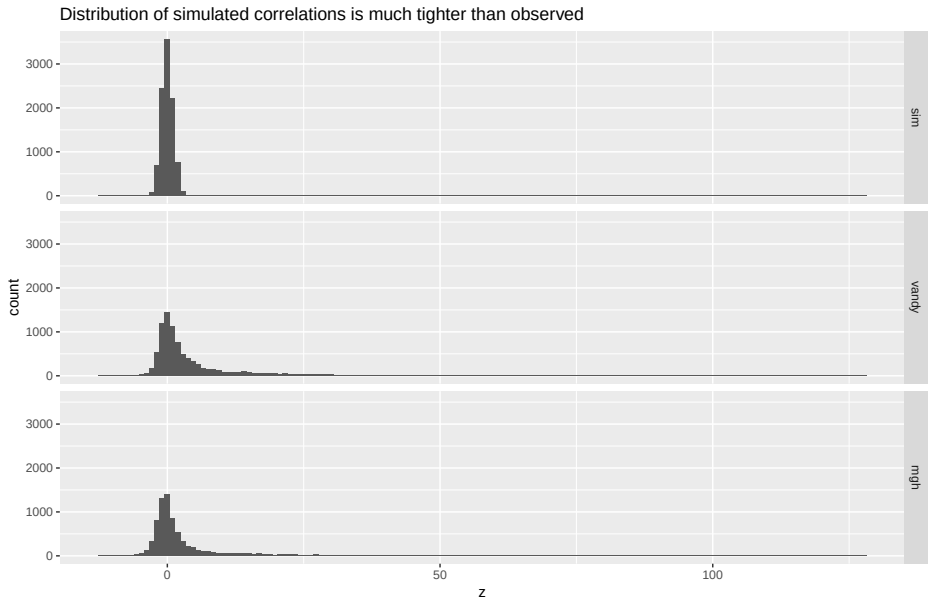

Figure 6.1: The distribution of association values from the simulated association network is much tighter around the null value of zero than those of either system.

6.2.2 Association correlations

Figure 6.2 shows a drastically decreased correlation between simulated association pairs and either system (plus combined). Correlations for the simulations are all positive but very small (e.g., 0.008 between the simulation and Vanderbilt's associations).

6.3 Discussion

These results show that the correlations between the two real systems are not driven exclusively, or even largely, by the inherent hierarchical structure of phecodes themselves, but by some other latent factors.

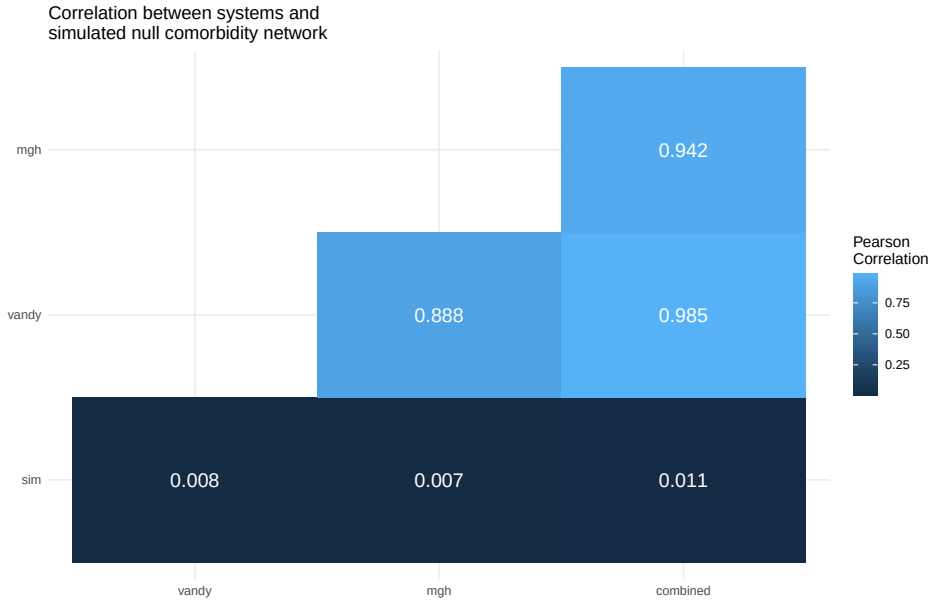

Figure 6.2: Correlations between association values from both systems along with simulated network. The observed system associations are much more correlated than we would expect just from the hierarchichal structure of the phecode's.

REFERENCES

(2020). National health expenditure accounts. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical. Accessed: 2020-9-21.

All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., and Dishman, E. (2019). The "all of us" research program. *N. Engl. J. Med.*, 381(7):668–676.

Avery, C. L., He, Q., North, K. E., Ambite, J. L., Boerwinkle, E., Fornage, M., Hindorff, L. A., Kooperberg, C., Meigs, J. B., Pankow, J. S., Pendergrass, S. A., Psaty, B. M., Ritchie, M. D., Rotter, J. I., Taylor, K. D., Wilkens, L. R., Heiss, G., and Lin, D. Y. (2011). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet.*, 7(10):e1002322.

Bauer, J., Ripperger, A., Frantz, S., Ergün, S., Schwedhelm, E., and Benndorf, R. A. (2014). Pathophysiology of isoprostanes in the cardiovascular system: implications of isoprostane-mediated thromboxane A2 receptor activation. *Br. J. Pharmacol.*, 171(13):3115–3131.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188.

Bojko, A. a. (2009). Informative or misleading? heatmaps deconstructed. In *Human-Computer Interaction. New Trends*, pages 30–39. Springer Berlin Heidelberg.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3: Data-Driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309.

Carroll, R. J., Bastarache, L., and Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, 30(16):2375–2376.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *Cityscape*, 1(2):1.

Chan, T. E., Stumpf, M. P. H., and Babtie, A. C. (2017). Gene regulatory network inference from Single-Cell data using multivariate information measures. *Cell Syst*, 5(3):251–267.e3.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R*. R package version 1.5.0.

Chen, Y. and Xu, R. (2014). Network analysis of human disease comorbidity patterns based on Large-Scale data mining. In *Bioinformatics Research and Applications*, pages 243–254. Springer International Publishing.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., and China Kadoorie Biobank (CKB) collaborative group (2011). China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.*, 40(6):1652–1666.

Chiquet, J., Rigaill, G., and Sundqvist, M. (2020). *aricode: Efficient Computations of Standard Clustering Comparison Measures*. R package version 1.0.0.

Cho, S. Y., Hong, E. J., Nam, J. M., Han, B., Chu, C., and Park, O. (2012). Opening of the national biobank of korea as the infrastructure of future biomedical science in korea. *Osong Public Health Res Perspect*, 3(3):177–184.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nat. Biotechnol.*, 29(11):987–991.

Csermely, P., Korcsmáros, T., Kiss, H. J. M., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.*, 138(3):333–408.

Davis, A., Gardner, B. B., and Gardner, M. R. (2009). *Deep South: A Social Anthropological Study of Caste and Class*. Univ of South Carolina Press.

De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *J. Classification*, 5(2):163–180.

de Leon, J., Dadvand, M., Canuso, C., White, A. O., Stanilla, J. K., and Simpson, G. M. (1995). Schizophrenia and smoking: An epidemiological survey in a state hospital. *Am. J. Psychiatry*, 152(3):453–455.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., and Crawford, D. C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6(290-297):18.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140.

Fisher, R. A. (1992). Statistical methods for research workers. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 66–70. Springer New York, New York, NY.

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., Przygodzki, R., and O'Leary, T. J. (2016). Million veteran program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.*, 70:214–223.

Gower, J. C. and Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Appl. Stat.*, 18(1):54.

Gustafsson, M., Hörnquist, M., and Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network–lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2(3):254–261.

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. *KDD*, 2015:387–396.

Harris, N. and Drton, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14(1):3365–3383.

Hebbring, S. J., Schrodi, S. J., Ye, Z., Zhou, Z., Page, D., and Brilliant, M. H. (2013). A PheWAS approach in studying HLA-DRB1*1501. *Genes Immun.*, 14(3):187–191.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, 106(23):9362–9367.

Hughes, J. R., Hatsukami, D. K., Mitchell, J. E., and Dahlgren, L. A. (1986). Prevalence of smoking among psychiatric outpatients. *Am. J. Psychiatry*, 143(8):993–997.

Kolaczyk, E. D. and Csárdi, G. (2020). *Statistical Analysis of Network Data with R*. Springer, Cham.

Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.-K., Kishore, N., Hao, T., Calderwood, M. A., Vidal, M., and Barabási, A.-L. (2019). Network-based prediction of protein interactions. *Nat. Commun.*, 10(1):1240.

Larremore, D. B., Clauset, A., and Jacobs, A. Z. (2014). Efficiently inferring community structure in bipartite networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 90(1):012805.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1983–1992.

Lohr, J. B. and Flynn, K. (1992). Smoking and schizophrenia. *Schizophr. Res.*, 8(2):93–102.

Luraschi, J. and Allaire, J. (2020). *r2d3: Interface to 'D3' Visualizations*. R package version 0.2.4.

LYON and B (2005). The opte project. *http://www.opte.org/*.

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., Struewing, J. P., Wolf, W. A., and eMERGE Team (2011). The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics*, 4:13.

Mcinnes, L. and Healy, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome Biol.*, 17(1):122.

Metsalu, T. and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res.*, 43(W1):W566–70.

Milanowski, L., Pordzik, J., Janicki, P. K., Kaplon-Cieslicka, A., Rosiak, M., Peller, M., Tyminska, A., Ozieranski, K., Filipiak, K. J., Opolski, G., Mirowska-Guzel, D., and Postula, M. (2017). New single-nucleotide polymorphisms associated with differences in platelet reactivity and their influence on survival in patients with type 2 diabetes treated with acetylsalicylic acid: an observational study. *Acta Diabetol.*, 54(4):343–351.

Minot, J. R., Arnold, M. V., Alshaabi, T., Danforth, C. M., and others (2020). Ratioing the president: An exploration of public engagement with obama and trump on twitter. *arXiv preprint arXiv*.

MOORE and F, E. (1959). The shortest path through a maze. *Proc. Int. Symp. Switching Theory, 1959*, pages 285–292.

Müller, S. M. (2016). *Wiring the World: The Social and Cultural Creation of Global Telegraph Networks*. Columbia University Press.

Newman, M. (2018). *Networks*. Oxford University Press.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U. S. A.*, 98(2):404–409.

Ng, C., Thubert, P., Watari, M., and Zhao, F. (2007). Network mobility route optimization problem statement. *Internet RFC4888, IETF*.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform.*, 10(4):354–366.

Pryke, A., Mostaghim, S., and Nazemi, A. (2007). Heatmap visualization of population based multi objective algorithms. In *Evolutionary Multi-Criterion Optimization*, pages 361–375. Springer Berlin Heidelberg.

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., Basford, M. A., Brown-Gentry, K., Balser, J. R., Masys, D. R., Haines, J. L., and Roden, D. M. (2010). Robust replication of Genotype-Phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.*, 86(4):560–572.

Schumacher, W. A., Steinbacher, T. E., Youssef, S., and Ogletree, M. L. (1992). Antiplatelet activity of the long-acting thromboxane receptor antagonist BMS 180,291 in monkeys. *Prostaglandins*, 44(5):389–397.

Sedgewick, R. and Wayne, K. (2011). *Algorithms*. Addison-Wesley Professional.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311.

Single, R. M., Strayer, N., Thomson, G., Paunic, V., Albrecht, M., and Maiers, M. (2016). Asymmetric linkage disequilibrium: Tools for assessing multiallelic LD. *Hum. Immunol.*, 77(3):288–294.

Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *Bull. Math. Biophys.*, 13(2):107–117.

Strayer, N., Ruderfer, D., and Xu, Y. (2020). Interactive network-based investigation of association matrices with associationsubgraphs. *ArXiv*, 1.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779.

Thomson, G. and Single, R. M. (2014). Conditional asymmetric linkage disequilibrium (ALD): extending the biallelic r2 measure. *Genetics*, 198(1):321–331.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(1):91–108.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. M. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, 14(5):535–542.

Wang, Y. and Zhu, J. (2014). Spectral methods for supervised topic models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 1511–1519. Curran Associates, Inc.

Werfel, T. A., Hicks, D., Rahman, B., Bendeman, W., Duvernay, M., Maeng, J. G., Hamm, H., Lavieri, R., Joly, M., Pulley, J., Elion, D., Brantley-Sieders, D., and Cook, R. (2020). Repurposing of a thromboxane receptor inhibitor based on a novel role in metastasis identified by phenome wide association study. *Mol. Cancer Ther.*

World Health Organization (1978). *International classification of diseases : [9th] ninth revision, basic tabulation list with alphabetic index*. World Health Organization.

World Health Organization (2004). *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. World Health Organization.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. (2018). Mining electronic health records (EHRs): A survey. *ACM Computing Surveys (CSUR)*, 50(6):85.

Yi, X., Lin, J., Zhou, Q., Huang, R., and Chai, Z. (2019). The TXA2R rs1131882, P2Y1 rs1371097 and GPIIIa rs2317676 three-loci interactions may increase the risk of carotid stenosis in patients with ischemic stroke. *BMC Neurol.*, 19(1):44.

Zheutlin, A. B., Dennis, J., Karlsson Linnér, R., Moscati, A., Restrepo, N., Straub, P., Ruderfer, D., Castro, V. M., Chen, C.-Y., Ge, T., Huckins, L. M., Charney, A., Kirchner, H. L., Stahl, E. A., Chabris, C. F., Davis, L. K., and Smoller, J. W. (2019). Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. *Am. J. Psychiatry*, 176(10):846–855.