



Systems biology

Discovering comorbid diseases using an inter-disease interactivity network based on biobank-scale PheWAS data

Yonghyun Nam ^{1,†}, Sang-Hyuk Jung ^{1,2,†}, Jae-Seung Yun ^{1,3}, Vivek Sriram ¹, Pankhuri Singhal ⁴, Marta Byrska-Bishop ⁵, Anurag Verma ⁶, Hyunjung Shin ⁷, Woong-Yang Park ⁸, Hong-Hee Won ^{2,8,*} and Dokyoon Kim ^{1,9,*}

¹Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, ²Department of Digital Health, SAIHST, Sungkyunkwan University, Samsung Medical Center, Seoul 06355, Republic of Korea, ³Division of Endocrinology and Metabolism, Department of Internal Medicine, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea, ⁴Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, ⁵New York Genome Center, New York, NY 10013, USA, ⁶Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, ⁷Department of Artificial Intelligence, Ajou University, Suwon 16499, Republic of Korea, ⁸Samsung Genome Institute, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, Republic of Korea and ⁹Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on May 2, 2022; revised on December 3, 2022; editorial decision on December 13, 2022; accepted on December 23, 2022

Abstract

Motivation: Understanding comorbidity is essential for disease prevention, treatment and prognosis. In particular, insight into which pairs of diseases are likely or unlikely to co-occur may help elucidate the potential relationships between complex diseases. Here, we introduce the use of an inter-disease interactivity network to discover/prioritize comorbidities. Specifically, we determine disease associations by accounting for the direction of effects of genetic components shared between diseases, and categorize those associations as synergistic or antagonistic. We further develop a comorbidity scoring algorithm to predict whether diseases are more or less likely to co-occur in the presence of a given index disease. This algorithm can handle networks that incorporate relationships with opposite signs.

Results: We finally investigate inter-disease associations among 427 phenotypes in UK Biobank PheWAS data and predict the priority of comorbid diseases. The predicted comorbidities were verified using the UK Biobank inpatient electronic health records. Our findings demonstrate that considering the interaction of phenotype associations might be helpful in better predicting comorbidity.

Availability and implementation: The source code and data of this study are available at <https://github.com/dokyoonkimlab/DiseaseInteractiveNetwork>.

Contact: wonhh@skku.edu or dokyoon.kim@penmedicine.upenn.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Comorbidity describes when a given patient has one or more additional medical conditions co-occurring alongside a primary disease (Lee *et al.*, 2008; Valderas *et al.*, 2009). Predicting comorbidity is

important for planning the clinical care of individual patients and for investigating clinical epidemiology; in particular, mortality risk is often increased when patients with underlying diseases are also diagnosed with adverse medical outcomes or complications (Cho *et al.*, 2021; Jørgensen *et al.*, 2012; Nashiry *et al.*, 2021).

Comorbidity studies involving multiple phenotypes at the population level can be divided into two categories according to the approach employed: (i) statistical analysis based on disease prevalence, or (ii) network analysis based on sharing of components among diseases. Notable studies using the statistical approach include Buddeke *et al.* (2019), a cohort study based on clinical records that presented complications with cardiovascular diseases (Buddeke *et al.*, 2019; Klimek *et al.* (2015), which used nationwide claim data to identify diabetes-related comorbidities (Klimek *et al.*, 2015); and Richardson *et al.* (2020), which revealed conditions comorbid with SARS-CoV-2 among 5700 hospitalized patients (Richardson *et al.*, 2020). Similarly, studies using network-based approaches and disease–disease associations include: Hidalgo *et al.* (2009), which built comorbidity maps for 657 diseases from the clinical records of over 30 million patients, with connectivity defined by prevalence-based comorbidity measures (Hidalgo *et al.*, 2009; Rubio-Perez *et al.* (2017), which explored potential comorbidity relationships by constructing a disease–disease network (DDN) based on shared disease-associated genes and their functions (Rubio-Perez *et al.*, 2017; Verma *et al.* (2019), which similarly identified disease comorbidities on the basis of shared genetic components by constructing a DDN based on a phenome-wide association study (PheWAS) that used electronic health record (EHR)-linked biobank data (Verma *et al.*, 2019; Dong *et al.*, 2021), which discovered disease multimorbidities using hospital inpatient data in the UK Biobank and provided biological explanations by constructing disease network with a set of genome-wide association studies (GWASs) data (Dong *et al.*, 2021).

Separately, several studies have attempted to categorize comorbidity as being direct or inverse (Ibáñez *et al.*, 2014; Sánchez-Valle *et al.*, 2020; Tabarés-Seisdedos and Valderas, 2013; Tarantino *et al.*, 2018); i.e. given an underlying disease, conditions that frequently co-occur at the population level are considered directly comorbid, while those with relatively few co-occurrences have inversely comorbid. Catalá-López *et al.* discovered that Alzheimer’s disease can decrease the co-occurrence of cancers from meta-analysis with literatures (Catalá-López *et al.*, 2014). Roitmann *et al.* revealed inverse comorbidity relationships from clustering analysis with health records in Demark (Roitmann *et al.*, 2014).

DDNs represent a map of topologies between phenotypes having shared biological components that may indicate phenotypic similarity and comorbidity. Several DDNs were constructed to provide insight into the human disease interactome by using genetic components shared with diseases, such as genes, proteins, or pathways (Barabási *et al.*, 2011; Goh *et al.*, 2007; Liu *et al.*, 2014; Zhou *et al.*, 2014). Other studies have used summary statistics obtained from GWASs to observe disease–disease associations by leveraging disease-associated single nucleotide polymorphisms (SNPs) (Darabos *et al.*, 2015; Dong *et al.*, 2021; Nam *et al.*, 2022; Verma *et al.*, 2019). However, these previous association-based DDN studies have not considered whether the associations or interactions that depend on shared components are agonistic or antagonistic. As shared genetic components can have different effects on different disease mechanisms, incorporating this directional information can improve the ability to predict comorbidity. For example, cardiovascular diseases have a synergistic association with low-density lipoprotein cholesterol (LDL-C) and an antagonistic association with high-density lipoprotein cholesterol (HDL-C): high LDL-C may increase the risk of cardiovascular disease, while high HDL-C may reduce it (Di Angelantonio *et al.*, 2009; Sharrett *et al.*, 2001). Insight into which disease pairs are likely or unlikely to co-occur may help in understanding the potential relationships of complex diseases. Moreover, as our knowledge of biological mechanisms advances, further genetic components of risk factors or protective factors related to disease mechanisms are revealed (von Mutius and Smits, 2020; Zhu *et al.*, 2018). Notably, Bulik-Sullivan *et al.* (2015) successfully identified negative genetic correlations that agreed with epidemiological associations (negatively correlated traits co-occur less than expected by chance), and also identified phenotypic correlations between complex traits and diseases via LD Hub and PhenoSpD as a follow-up study (Bulik-Sullivan *et al.*, 2015; Zheng *et al.*, 2017, 2018). They utilized linkage disequilibrium score

(LDSC) regression to estimate the strength of genetic correlations between diseases by comparing association test statistics between two diseases according to LD score trait in a polygenic model. Therefore, LDSC-based methods effectively calculate genetic correlations mainly for polygenic traits, using all variants regardless of their statistical significance in GWAS (Bulik-Sullivan *et al.*, 2015). Because quantifying genetic correlations using LDSC regression focused on the entire spectrum of variants in GWASs, it is hard to directly apply their method to estimate disease interactions based on knowledge-based shared genetic components (e.g. disease–gene associations cataloged in the Online Mendelian Inheritance in Man database) or data-driven shared components with statistical thresholds (e.g. based on significant variants in genome-wide data).

The association-based DDNs that consider only the total amount of shared components are not easy to determine whether the relationships between diseases are agonistic or antagonistic. In contrast, the correlation-based DDNs computed by LDSC regression can determine disease interactions but it may include more false-positive disease associations in the resulting networks than those considering only significant SNPs. This motivates us to explore interactions between diseases considering the direction of effect of shared genetic components that reach genome-wide significance in PheWAS summary statistics and to study how inter-disease interactivity is related at the epidemic and genomic levels.

Here, we propose a novel network-based framework to build an inter-disease interaction network that can take into account the direction of effects for components shared between diseases. We consider phenotypes as having either synergistic or antagonistic association: synergistic if two diseases mutually increase the tendency of their co-occurrence, or antagonistic if they decrease the tendency of co-occurrence. To examine whether the interactivity of disease associations can help predict/prioritize comorbidity, we constructed a DDN that considers the direction of effect of SNPs significantly associated with diseases, obtained from PheWAS summary statistics. This inter-disease interactivity network is a signed disease–disease network (signed DDN) having both positive and negative edges in the graph (i.e. positive values for a synergistic association and negative for an antagonistic association). To translate synergistic and antagonistic association at the genomic level to direct and inverse comorbidity at an epidemic level, we also develop a novel graph-based semi-supervised learning (SSL) for comorbidity scoring, modifying the objective functions of label propagation algorithms to work for a signed DDN with a unary label by introducing a signed degree matrix and signed graph Laplacian (Gallier, 2016). The scoring algorithm prioritizes comorbid diseases in relation to an index disease of interest.

2 Materials and methods

The overall procedure consists of three main parts (Fig. 1): (i) constructing the signed DDN based on shared genetic components, (ii) predicting comorbidity scores using graph-based SSL and (iii) prioritizing/ranking comorbid conditions in relation to a given disease.

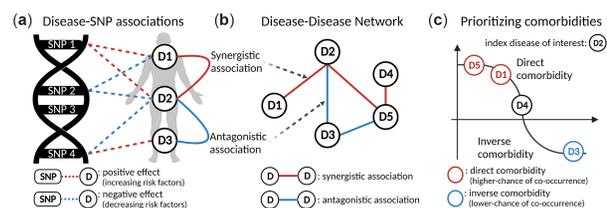


Fig. 1. Schematic description of the proposed method. (a) Defining directionality of disease–SNP associations. D1 and D2 have a synergistic association because the overall effects of their shared SNPs have consistent direction, whereas D2 and D3 have an antagonistic association because the effects of their shared SNPs have opposite direction. (b) DDN incorporating synergistic and antagonistic associations. We constructed a signed DDN with pre-defined/pre-calculated disease associations. (c) Prioritizing comorbidity. Higher scores mean higher chance of disease co-occurrence; D5 has the highest chance of comorbidity with D2

First, we built a signed DDN based on the sharing of significant SNPs between disease pairs, which relied on data obtained from UK Biobank PheWAS summary statistics. We determined disease associations by accounting for the overall direction of effects of genetic components shared between diseases, and categorize those associations as synergistic or antagonistic. In this network, two diseases have a synergistic association if their shared SNPs have the overall same direction of effect, or an antagonistic association if the SNP effects are in opposition. For example, Disease 1 (D1) and Disease 2 (D2) have shared components in SNP1 and SNP2, whereas D2 and D3 share SNP4 (Fig. 1a). Both SNP1 and SNP2 display the same direction of effect on D1 and on D2: SNP1 is associated with positive direction of effects (increasing risk), while SNP2 is associated with negative direction of effects (decreasing risk). Thus, the association between D1 and D2 is synergistic. In contrast, D2 and D3 have an antagonistic association because the respective effect of SNP4 is opposite (i.e. positive for D2 but negative for D3). Once the associations are determined, a signed DDN is constructed to represent the relationships between phenotype pairs as a graph with nodes and edges (Fig. 1b).

Second, we perform comorbidity score prediction: given a specific disease, we predict comorbidity of other diseases by applying graph-based SSL to the constructed network (Lee *et al.*, 2020; Nam *et al.*, 2019a,b). Since the network includes both synergistic and antagonistic associations (edge weights with positive and negative values), we posited the following hypotheses concerning comorbidity predictions: (i) two diseases have a chance of comorbidity if they are connected (i.e. have at least one shared SNP), regardless of association direction; (ii) if two diseases have a synergistic association according to overall direction of shared SNPs, they are more likely to co-occur (high chance of co-occurrence); and (iii) if two diseases have an antagonistic association, they are less likely to co-occur (low chance of co-occurrence). If these hypotheses are supported, scoring algorithms can predict the comorbidity of another medical condition given a specific underlying disease.

Finally, we prioritize the chance of disease co-occurrence based on the predicted comorbidity scores, stratifying by deciles. All diseases in a network have some chance of comorbidity with the index disease, but the degree of association can range considerably; a representative example is illustrated in Figure 1c. This ranking is then validated using disease prevalence-based co-occurrence measures estimated from EHRs (Hidalgo *et al.*, 2009).

2.1 Constructing the DDN from PheWAS summary data

The proposed DDN was built with a shared genetic component hypothesis, namely that two different phenotypes were linked if they shared the significant SNPs from the PheWAS summary statistics. We obtained the beta-coefficient (β_{ik}) and standard error (SE_{ik}) values for the association between phenotype i and SNP k when that association passed the P -value threshold for significance in summary statistics. Then, we selected an appropriate significance threshold (P -value $< 1 \times 10^{-4}$) and constructed a disease-SNP association matrix: specifically, given m phenotypes and k SNPs, we constructed the association matrix $R \in \mathbb{R}^{m \times k}$ (Fig. 2a). For each SNP k associated with disease i , the element $r_{ik} = \beta_{ik}/SE_{ik}$ is a constant value of z -score consisting of the corresponding beta-coefficient divided by the standard error. The sign of r_{ik} indicates the direction of effect of SNP k in relation to disease i . The DDN was then constructed from this matrix to represent the genetic associations between pairs of phenotypes by measuring proximity between disease-SNP vectors.

This DDN is an undirected, signed and weighted graph $G = (V, W)$ in which $V = \{v_i \in \mathbb{R}^k \mid i = 1, 2, \dots, m\}$ denotes the set of diseases and $W = \{w_{ij} \mid i, j = 1, 2, \dots, m\}$ denotes the similarity between diseases. Each disease $v_i = \{r_{i1}, r_{i2}, \dots, r_{ik}\}$ has k -dimensional SNP vectors, and the similarity w_{ij} between any two diseases v_i and v_j is calculated as follows:

$$w_{ij} = \frac{\sum_{k \in S} (r_{ik}) \cdot (r_{jk})}{\sqrt{\sum_{k=1}^K r_{ik}^2} \sqrt{\sum_{k=1}^K r_{jk}^2}} = \frac{\sum_{k \in S} (\beta_{ik}/SE_{ik}) \cdot (\beta_{jk}/SE_{jk})}{\sqrt{\sum_{k=1}^K (\beta_{ik}/SE_{ik})^2} \sqrt{\sum_{k=1}^K (\beta_{jk}/SE_{jk})^2}}, \quad (1)$$

where r_{ik} and r_{jk} are the respective z -scores associated with SNP k for diseases v_i and v_j ; K is the total number of SNPs; and S is the set of significant SNPs shared between v_i and v_j . Since the z -score r_i can take a negative value, each edge weight w_{ij} can range from -1 to 1 . Figure 2a illustrates the constructed disease similarity matrix (W) with both positive and negative associations. The magnitude of a weight ($|w_{ij}|$) between a disease pair describes the number of significant SNPs those diseases share. A synergistic association ($w_{ij} > 0$) means that two diseases have shared SNPs and the overall effect of those SNPs has the same direction for both diseases. An antagonistic association ($w_{ij} < 0$) indicates that the shared SNPs have overall opposite directions of effect for the two diseases. The similarity matrix can quantify how similar two diseases are in terms of the directions of SNP effects.

2.2 Network-based comorbidity scoring algorithms

Once the DDN incorporating synergistic and antagonistic associations (signed DDN) was constructed, comorbidity scoring was performed using a graph-based SSL, which is employed for scoring algorithms (Chong *et al.*, 2020; Subramanya and Talukdar, 2014). Figure 2b describes the problem setup for predicting comorbidity scores. The network-based scoring algorithm used here is a transductive learning approach (Chong *et al.*, 2020); it predicts the co-occurrences between an underlying disease and other diseases when only the underlying disease is known to occur (one positive sample), and all others have unknown occurrences (unlabeled samples). Those scores can provide prioritized comorbidity scores for the unlabeled diseases by propagating with disease associations from the signed DDN. The proposed network simultaneously contains both positive and negative edges, obtained from (1); i.e. two diseases having SNPs in common can co-occur regardless of the respective directionality of SNP effects. Nevertheless, the chance of co-occurrence for two synergistically associated diseases may be relatively higher than that of two diseases having an antagonistic association, since the former pair shares many SNPs with the same direction of effect while the latter features SNPs with opposite directions of effect.

The following describes the procedure of comorbidity scoring. Consider a signed DDN $G = (V, W)$, where the similarity matrix W has both positive and negative edge weights. Let $y = (y_1, \dots, y_m)^T$ denotes the initial labels for the set of diseases and $f = (f_1, \dots, f_m)^T$ the set of predicted comorbidity scores. We set a unary label ($y_l = +1$) for an index disease of interest (v_l) and set unlabeled ($y_l = \{0\}$) for remaining other diseases. Label information on the labeled disease is propagated to unlabeled diseases on graph G to obtain real-valued scores f , with two assumptions: the *smoothness condition* (predicted scores f_i should not be too different from the f_j values in adjacent unlabeled nodes) and the *loss condition* (predicted scores f_i should be close to the given label of y_i). Generally, for an unsigned graph, $\tilde{G} = (V, \tilde{W})$, the smoothness condition is represented as

$$f^T L f = \sum_{i \sim j} \tilde{w}_{ij} (f_i - f_j)^2, \quad (2)$$

where the graph Laplacian L is defined as $L = D - \tilde{W}$, \tilde{W} being the unsigned similarity matrix, $D = \text{diag}(d_i)$ the degree matrix and $d_i = \sum_j \tilde{w}_{ij}$. However, since the proposed network has signed edges from (1), we modified the smoothness condition of (2) to handle both positive and negative associations by incorporating a signed degree matrix:

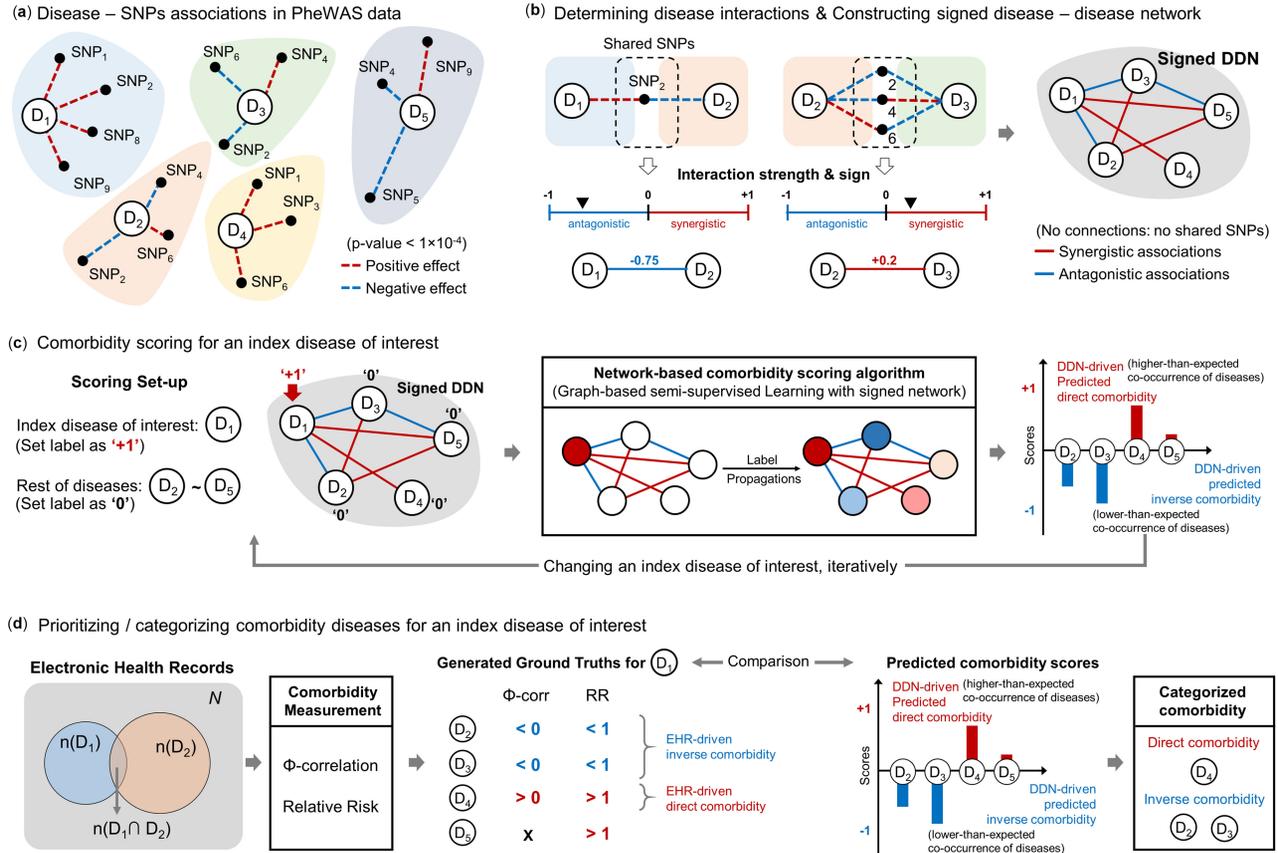


Fig. 2. Step-by-step explanation of the proposed method. (a) Visualization of disease-SNPs associations in PheWAS summary data. The positive/negative direction of effect of SNPs (dashed line) depicts the direction of beta-coefficients obtained from summary statistics. (b) Constructing inter-disease interactive network (signed DDN) with determining synergistic and antagonistic associations. (c) Semantic description of proposed comorbidity scoring algorithm with graph-based SSL. (d) Comparison scheme with predicted DDN-driven comorbidity and EHR-driven comorbidity. Based on this explanation, we provide pseudocode in [Supplementary Algorithm S1](#)

$$f^T \bar{L} f = \sum_{i \sim j} |w_{ij}| (f_i - \text{sign}(w_{ij}) f_j)^2, \quad (3)$$

where \bar{L} is the signed graph Laplacian, defined as $\bar{L} = \bar{D} - W$, in which $\bar{D} = \text{diag}(\bar{d}_i)$ is the signed degree matrix with $d_i^- = \sum_j |w_{ij}|$ and $\text{sign}(\cdot)$ is the signum function. The signed graph Laplacian \bar{L} is positive semidefinite, and the quantity of $f^T \bar{L} f$ is non-negative (Gallier, 2016). The loss condition is $(f - y)^T (f - y) = \sum_i (f_i - y_i)^2$. The predicted output f is obtained by minimizing the following quadratic function using the loss and smoothness conditions:

$$\min_f (f - y)^T (f - y) + \mu f^T \bar{L} f. \quad (4)$$

The closed form solution is obtained as

$$f = (I + \mu \bar{L})^{-1} y, \quad (5)$$

where the hyper-parameter μ trades off loss and smoothness. The resulting predicted outcome f on unlabeled nodes has either positive or negative values. We denote the predicted comorbidity scores f as the DDN-driven comorbidity, which indicates the relative likelihood of two diseases co-occurring based on the direction of effects of genetically associated factors.

2.3 Prioritizing and categorizing comorbidity

We prioritize and categorize the list of co-occurring diseases as having higher or lower chance of co-occurrence. Hereafter, we denote the higher chance of comorbidity as *direct comorbidity* and the lower chance as *inverse comorbidity* (Catalá-López et al., 2014;

Tabarés-Seisdedos and Baudot, 2016). We categorize DDN-driven comorbidity f into direct and inverse comorbidity. For validating DDN-driven comorbidity, we used co-occurrence measures estimated from EHR (denoted as EHR-driven comorbidity). EHR-driven comorbidity is defined in terms of the disease prevalence-based ϕ -correlation and relative risk (Hidalgo et al., 2009).

The Pearson correlations for binary variables (ϕ -correlation) and relative risk were used to estimate EHR-driven comorbidity (Fig. 2c). Given N patients and the list of m diagnoses in clinical records, we constructed a $(\frac{m^2 - m}{2})$ binary contingency matrix for each disease pair to display the frequency distribution. The ϕ -correlation for the pair of diseases i and j is calculated by $\phi_{ij} = \frac{C_{ij}N - P_i P_j}{\sqrt{P_i P_j (N - P_i)(N - P_j)}}$ based on their prevalence, where C_{ij} is the number of patients diagnosed with both diseases while P_i and P_j indicate the number of patients diagnosed with disease i and j , respectively. A positive correlation means two diseases tend to occur simultaneously and a negative correlation that they tend not to co-occur. We also captured tendency of co-occurrence among disease pairs by calculating relative risks. The relative risk for two diseases $RR_{ij} = \frac{C_{ij}N}{P_i P_j}$ is measured as the ratio of risk between the diseases based on the disease prevalence. $RR_{ij} > 1$ indicates that both diseases co-occur more frequently than expected by chance, and $RR_{ij} < 1$ indicates their less frequent co-occurrence than expected by chance (Hidalgo et al., 2009; Sánchez-Valle et al., 2020). A disease pair i and j was defined as having EHR-driven comorbidity when both measures indicated a tendency to co-occur ($\phi_{ij} > 0$ and $RR_{ij} > 1$).

Taking as ground truths the corresponding EHR-driven comorbidity, DDN-driven comorbidity scores f from Equation (5) for each unlabeled disease can be decomposed into f^+ and f^- based on a

threshold value. Suppose that $\hat{y} = (\hat{y}_1, \dots, \hat{y}_m)^T$ is the ground truths, then \hat{y}_i can take a value of ‘+1’ if disease i is defined as having EHR-driven comorbidity, and ‘-1’ otherwise. Then, the receiver operating characteristic (ROC) curve was analyzed to determine a threshold value J for deciding the labels on f . Youden’s J statistic was used for a threshold value where the maximum value of $J = \text{sensitivity} + \text{specificity} - 1$, as in a ROC curve (Schisterman *et al.*, 2005). We categorized diseases as having DDN-driven direct comorbidity (high chance of co-occurrence) when $f^+ = \{f_i | f_i \geq J, \text{ for } \forall i\}$ and DDN-driven inverse comorbidity (low chance of co-occurrence) when $f^- = \{f_i | f_i < J, \text{ for } \forall i\}$.

3 Results

3.1 Data processing for network construction and validation

Disease-SNP associations were generated from UK Biobank PheWAS summary statistics based on EHR-derived broad phenotype codes (PheCode) (Wei *et al.*, 2017; Wu *et al.*, 2019). The PheWAS summary statistics were obtained from <https://www.lee.labs.org/resources>, and include 1403 phenotypes (Zhou *et al.*, 2018), of which 976 were excluded in this study due to having a relatively low number of cases (<1000) or representing injuries and poisonings, symptoms, or sex-specific disease. Ultimately, network construction and comorbidity prediction were performed on 427 diseases belonging to 14 phenotype categories (Fig. 3). Pre-processing of the PheWAS summary results was performed using the following steps: (i) 588 711 independent SNPs were selected by linkage disequilibrium pruning with a threshold (window size: 50 kb, step size: 5 kb, and r^2 threshold: 0.5), and (ii) SNPs with association P -value $< 1 \times 10^{-4}$ were selected, yielding 39 382 SNPs for analysis. Disease-SNP associations were represented as a $(427 \times 39\,382)$ -dimensional matrix in which the association of each element pair was represented by a z -score. This matrix was then used to construct a DDN with signed associations using Equation (1).

EHR-driven comorbidity information was collected and generated from the UK Biobank hospital episode statistic database. Disease diagnoses for 502 505 UK Biobank participants are represented by ICD-9 and ICD-10 codes, which were mapped to PheCodes by referring to PheCode Map versions 1.2 and 1.2b1 (<https://phewascatalog.org/>). A patient-phenotype matrix was constructed with binary values indicating the diagnosis status of each participant for 427 diseases to calculate ϕ -correlation and relative risk. More details about the UK Biobank PheWAS results are provided in Supplementary Text S1, and the phenotypes used in this analysis are listed in Supplementary Data S1.

3.2 Construction of the inter-disease interactivity network (signed DDN)

The resulting inter-disease interactivity network (Fig. 3a) consisted of 427 phenotypes (nodes) and 9223 total associations (edges), of which 6209 were synergistic (67%; red) and 3014 antagonistic (33%; blue) (Fig. 3c and Supplementary Figs S1 and S2). We examined the distribution of associations across the 14 phenotype categories (Fig. 3b), and found connections between diseases within a phenotype category to typically be synergistic. Namely, diseases belonging to a given phenotype category often had consistent effect directions for their shared SNPs, which means that the chance of their co-occurrence may be higher. One feature of signed DDNs is the ability to observe genetic associations between diseases as the direction of effect of shared SNPs. Our signed DDN was constructed while taking into account both the total amount and direction of shared SNPs. The greater the number of shared SNPs, the stronger the association between two diseases; likewise, the more homogeneous the direction of effect, the stronger the synergistic association. We examined the composition of synergistic/antagonistic associations in our signed DDN. Of the 427 diseases in the network, 364 feature both associations, 57 have only synergistic associations, 3 have only antagonistic associations and, remaining 3 diseases have

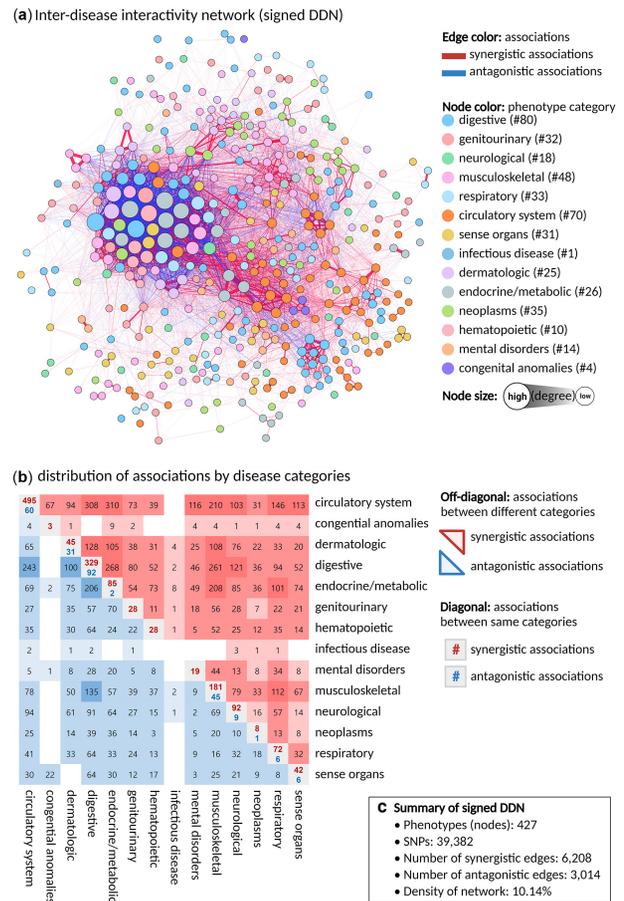


Fig. 3. Visualization of an inter-disease interactivity network (signed DDN). (a) A signed DDN with 427 phenotypes. Nodes represent diseases, and edges represent the shared SNP associations between pairs of diseases. Node color indicates phenotype category, and size the degree of connections. Each line indicates either synergistic (positive values, Supplementary Fig. S1) or antagonistic (negative values, Supplementary Fig. S2) associations between nodes. Supplementary Figure S3 shows heatmap for signed DDN (b) distribution of disease associations among the 14 phenotype categories. Upper and lower diagonal elements, respectively, depict the numbers of positive and negative associations across different categories. The main diagonal represents associations among diseases belonging to the same phenotype category. (c) Summarized information on the data used for constructing the signed DDN

no associations (Supplementary Data S2, <https://hdp.biomedinfo.lab.com/ddn/signedDDN>). To investigate a deeper biological interpretations of the network, we described the composition of phenotypes and interactions in signed DDN for coronary atherosclerosis (PheCode: 411.4) as case example (Supplementary Text S2). For example, Coronary atherosclerosis and type 2 diabetes (PheCode: 250.2) shared 66 SNPs, of which 55 SNPs were associated with positive direction of effects and 11 with negative direction of effects in relation to both diseases. By calculating the overall direction of effects (similarity = 0.29), coronary atherosclerosis and type 2 diabetes were determined to have a synergistic association.

3.3 Comparison of comorbidity prediction performance

The objective of network-based comorbidity scoring algorithms using a signed DDN is to predict whether other diseases are more or less likely to co-occur when one disease occurs. We hypothesized that the direction of effects of shared SNP is informative in predicting comorbidity and in prioritizing disease co-occurrence. Namely, higher scores indicate greater sharing of SNPs and more consistent direction of SNP effects; thus, diseases having stronger synergistic association with a given index disease have higher predicted

comorbidity scores. To test this hypothesis, we compared prediction performance between the proposed method (signed DDN) and the baseline method (unsigned DDN). The unsigned DDN was constructed as in previous studies (Goh et al., 2007; Nam et al., 2019a; Verma et al., 2019), with only the number of shared genetic components being considered. The difference between signed and unsigned DDNs is explained in Supplementary Figure S4. A transductive approach was used, in which only one set of experiments was carried per disease. EHR-driven comorbidity was used for ground truths. Prediction performance was evaluated in terms of the area under the receiver operating characteristics curve (AUC) and Spearman rank correlation (ρ) (Fawcett, 2006; Yule, 1919). Both AUC and rank correlation for an index disease of interest were obtained by comparing the predicted DDN-driven comorbidity scores and EHR-driven comorbidity scores (Supplementary Fig. S5).

Comparison of prediction performance between the signed and unsigned DDNs revealed that the interaction of the association between phenotypes can be helpful in predicting comorbidity (Fig. 4). The AUC distribution of the 427 phenotypes in the signed DDN was shifted higher relative to that of the unsigned DDN (Fig. 4a). Moreover, the signed DDN demonstrated superior performance, achieving an average of 0.601 (versus 0.573 with the unsigned DDN) and with values generally falling above the diagonal when plotted against corresponding unsigned scores (Fig. 4b). In the same manner, we compared Spearman rank correlation coefficients between the two DDNs (Fig. 4c). In the signed DDN, 341 diseases had positive correlations (above zero on y-axis, Fig. 4c), of which 259 exhibited significant correlations (0.103–0.556, P -value < 0.05). In the unsigned DDN, 163 diseases had positive correlations (above zero on x-axis, Fig. 4c), of which 53 were statistically significant (0.105–0.290, P -value < 0.05). Only five diseases statistically significant in the unsigned DDN achieved higher positive correlation coefficients than were determined with the signed DDN. Thus, direct comparison of DDN-comorbidity score and prevalence-based relative risk indicates that taking into account the direction of effect of shared SNPs could better predict disease comorbidity. Notably,

when not considering direction of effect, only weakly positive or non-significant correlations were obtained between prevalence-based relative risk and predicted scores. Both AUCs and Spearman rank correlation coefficients can be interpreted as the explanatory power of comorbidity predictions for index diseases. Performance measurement can be affected by the statistical power of PheWAS for each disease, but they are affected more greatly by the structure of the network because they are the predicted result through a scoring algorithm. Thus, an index disease with a high AUC means that disease associations explained by genetic information can be interpreted as comorbidity relationships in epidemiology.

3.4 Interpretation of the correlation coefficient in signed/unsigned DDNs

We further investigated the top 15 diseases with significant rank correlations (marked as red in Fig. 4c). DDN-comorbidity scores for these diseases in the signed DDN were highly correlated with prevalence-based relative risk ($\rho > 0.5$ with P -value < 0.05), whereas those in the unsigned DDN were relatively low ($0 < \rho < 0.3$). Overall, the top 15 included 11 circulatory-related diseases that had high positive correlation coefficients in the signed DDN (Fig. 5a). Four diseases with highly significant correlations in the signed DDN did not achieve significance in the unsigned DDN (superscript in P -value, Fig. 5a). It can be inferred that considering SNP direction of effect and its commonality between diseases is capable of capturing complex disease associations that might not be identified when only considering sharing of SNPs.

In this approach, synergistically associated diseases directly connected to the index disease are likely to have the highest predicted scores. Consequently, we examined which diseases had direct connections when taking each of the 15 top diseases as the index disease (Fig. 5b). The results showed all diseases with high correlation coefficients to have relatively high ratios of synergistic association. In addition, we found that comorbidity scores derived from the signed DDN were highly correlated with EHR-driven comorbidity, even for diseases having many antagonistic associations, such as hypercholesterolemia, disorders of lipid metabolism, myocardial infarction and coronary atherosclerosis. This further supports that considering the direction of effects is helpful when using PheWAS data to analyze the complex relationships between diseases, as the latent associations between diseases (synergistic/antagonistic association) provide important information for comorbidity prediction.

3.5 Clinical implication of comorbidity scoring based on the signed DDN

We demonstrated how to predict disease comorbidity by applying a graph-based SSL for signed graph to a signed DDN encompassing 427 diseases. To provide a case study illustrating the proposed method and its interpretation, coronary atherosclerosis was selected from among the top 15 diseases having significant rank correlations in the signed DDN. Coronary atherosclerosis is a frequent cause of coronary artery disease, a common chronic condition for which disease risk is characterized by a substantial and complex polygenic contribution, with heritability between 40% and 60% (Fischer et al., 2005; Marenberg et al., 1994; McPherson and Tybjaerg-Hansen, 2016). We provided the comorbidity scores for a list of direct comorbidity diseases (Table 1), and the comorbidity score curve obtained from Equation (5) (Fig. 6a). To suggest the practical use of comorbidity scores, we divided the scores into deciles and categorized accordingly, ranging from high risk of comorbidity (direct comorbidity) to low risk of comorbidity (inverse comorbidity). We decided the Tier-1 diseases as direct comorbidity group and Tier-10 diseases as inverse comorbidity group according to their ranking of predicted scores. The remaining diseases between Tier-2 and Tier-9 were relatively lower co-occurrence than Tier-1 or higher co-occurrence than Tier-10.

We examined the sub-networks for diseases (Fig. 6b) having high risk of co-occurrence (Tier-1) and those having low risk of co-occurrence (Tier-10). All Tier-1 diseases were synergistically associated with coronary atherosclerosis and also with each other, which

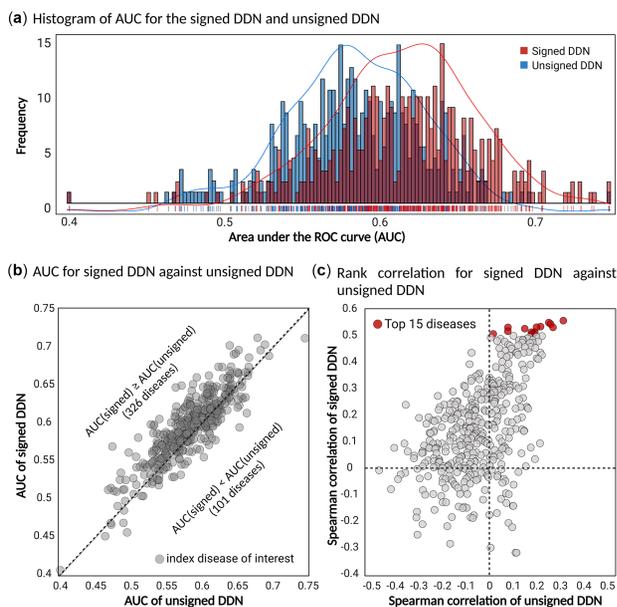


Fig. 4. Improvement of performance with the proposed signed DDN in comparison to an unsigned network. (a) Histogram of AUC values for the signed and unsigned DDNs. AUC values are binned along the x-axis. The bars depict the disease frequency in each bin for both networks. The histogram is shifted towards right for the signed DDN. (b) Plot of signed DDN AUC against that of the unsigned DDN. The y- and x-axis indicate AUC values for the signed and unsigned networks, respectively. Each circle indicates a set of experiments for one disease. (c) Comparison of Spearman rank correlations. Spearman rank correlations indicate ordinal associations between rankings of DDN-comorbidity scores (f) and rankings of EHR-driven relative risks

(a) List of the top 15 diseases with significant rank correlation in the signed DDN

| PheCode | Phenotypes | Signed DDN | | | Unsigned DDN | | | Spearman rank correlation coefficient (ρ) | | | | | | | | |
|---------|--|------------|------------------|------------------------|--------------|-------------------|-----------------------|--|---|-----|-----|-----|-----|-----|-----|-----|
| | | ρ | 95% CI | p-value | ρ | 95% CI | p-value | -0.1 | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 |
| 747.13 | Congenital anomalies of great vessels | 0.5563 | [0.4785, 0.6253] | 2.98x10 ⁻²⁹ | 0.2903 | [0.1902, 0.3844] | 4.37x10 ⁻⁸ | | | | | | | | | |
| 426.32 | Left bundle branch block | 0.5490 | [0.4714, 0.6181] | 5.10x10 ⁻²⁹ | 0.2324 | [0.1309, 0.3290] | 1.09x10 ⁻⁵ | | | | | | | | | |
| 426.91 | Cardiac pacemaker in situ | 0.5435 | [0.4670, 0.6119] | 1.64x10 ⁻²⁹ | 0.2393 | [0.1403, 0.3337] | 3.64x10 ⁻⁶ | | | | | | | | | |
| 426.31 | Right bundle branch block | 0.5339 | [0.4546, 0.6047] | 2.99x10 ⁻²⁷ | 0.1995 | [0.0969, 0.2980] | 1.68x10 ⁻⁴ | | | | | | | | | |
| 411.1 | Unstable angina | 0.5316 | [0.4544, 0.6009] | 1.89x10 ⁻²⁸ | 0.1836 | [0.0834, 0.2802] | 3.78x10 ⁻⁴ | | | | | | | | | |
| 272.11 | Hypercholesterolemia | 0.5309 | [0.4573, 0.5973] | 5.69x10 ⁻³¹ | 0.0696 | [-0.0278, 0.1657] | 0.1612* | | | | | | | | | |
| 394.3 | Aortic valve disease | 0.5308 | [0.4505, 0.6026] | 1.80x10 ⁻²⁶ | 0.2493 | [0.1476, 0.3458] | 2.75x10 ⁻⁶ | | | | | | | | | |
| 433.1 | Occlusion and stenosis of precerebral arteries | 0.5261 | [0.4411, 0.6018] | 9.40x10 ⁻²⁴ | 0.1364 | [0.0261, 0.2434] | 1.55x10 ⁻² | | | | | | | | | |
| 272 | Disorders of lipid metabolism | 0.5167 | [0.4419, 0.5844] | 2.71x10 ⁻²⁹ | 0.0691 | [-0.0281, 0.1650] | 0.1631* | | | | | | | | | |
| 411.2 | Myocardial infarction | 0.5130 | [0.4339, 0.5843] | 2.69x10 ⁻²⁶ | 0.1765 | [0.0760, 0.2734] | 6.39x10 ⁻⁴ | | | | | | | | | |
| 411.4 | Coronary atherosclerosis | 0.5129 | [0.4345, 0.5835] | 9.60x10 ⁻²⁷ | 0.1631 | [0.0632, 0.2597] | 1.47x10 ⁻³ | | | | | | | | | |
| 426.3 | Bundle branch block | 0.5065 | [0.4111, 0.5910] | 1.12x10 ⁻¹⁸ | 0.1717 | [0.0523, 0.2863] | 5.07x10 ⁻³ | | | | | | | | | |
| 425.1 | Intrinsic cardiomyopathies | 0.5064 | [0.4165, 0.5864] | 9.57x10 ⁻²¹ | 0.0109 | [-0.1030, 0.1246] | 0.8512* | | | | | | | | | |
| 414 | Chronic heart disease | 0.5057 | [0.4243, 0.5790] | 1.22x10 ⁻²⁴ | 0.1643 | [0.0617, 0.2635] | 1.81x10 ⁻³ | | | | | | | | | |
| 443.9 | Peripheral vascular disease | 0.4994 | [0.4196, 0.5716] | 3.56x10 ⁻²⁵ | 0.0915 | [-0.0096, 0.1908] | 0.0759* | | | | | | | | | |

(b) Composition of directly connected diseases with an index disease of interest in signed DDN

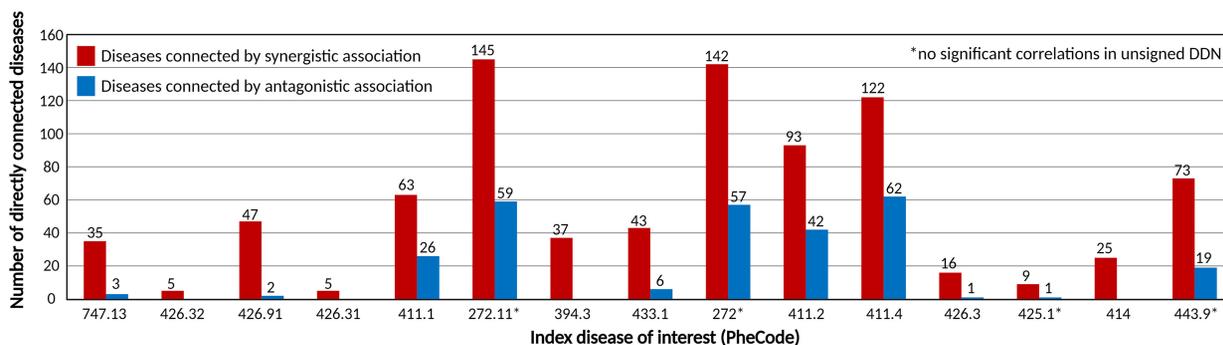


Fig. 5. List of the top 15 diseases with significant rank correlations in the signed DDN. (a) Spearman rank correlation in the signed DDN against the unsigned DDN. Diseases were sorted in descending order with respect to correlation coefficient in signed DDN. (b) Composition of directly connected diseases. Red bars indicate the number of synergistic associations and blue bars the number of antagonistic associations among diseases directly connected with the given index disease. Four diseases with a superscript in *P*-value were not statistically significant in the unsigned DDN

means that high-ranked diseases overall have shared SNPs with consistent direction of effect. In contrast, within the Tier-10 group, all diseases directly connected with coronary atherosclerosis had antagonistic associations and the predicted scores were negative. In Tier-10 group, macular degeneration (PheCode: 362.29) was antagonistically associated with coronary atherosclerosis and the predicted comorbidity was inverse. The scoring algorithm with the signed DDN detected the known relationships that high levels of HDL-C were associated with low risk of coronary artery disease and associated with high risk of age-related macular degeneration (Fan *et al.*, 2017). A sub-network consisting of all diseases directly connected with coronary atherosclerosis is given in Supplementary Figure S6. The top 20 diseases having direct comorbidity with coronary atherosclerosis are summarized in Table 1.

3.6 Criterion of leveraging disease-SNP associations in summary statistics

We developed a signed DDN by leveraging the overall direction of the effects of genetic components shared between diseases. We also performed predictions of co-occurrence diseases given an index disease to examine whether the signed DDN can have more reliable explanatory power for disease interactions. To propagate the seed

label information of index disease to the rest of the diseases in a network through the scoring algorithm, we took a less stringent genome-wide significance level (P -value $< 1 \times 10^{-4}$) for network construction. However, depending on the arbitrary selection of significance level, the underlying structure of the network can be changed. The disease interactions are sparser at the most stringent level and denser at the less stringent level. It is necessary to explore which P -value threshold in the disease-SNP associations for defining disease associations. We built another comparative DDN by performing LDSC regression to investigate whether comorbidity prediction results varied according to the number of SNPs in building a network. We built the genetic correlation-based DDN (denoted as LDSC-DDN) by performing LDSC regression with UK Biobank PheWAS summary data to estimate the genetic correlations of 427 pairs used for the signed DDN (Bulik-Sullivan *et al.*, 2015). The LD scores were calculated from European samples in the 1000 Genomes Project phase 3 database (Auton *et al.*, 2015). We considered disease pairs with positive and negative correlation values with significance (P -value < 0.05). We took the correlation matrix as the similarity matrix of the DDN and applied the proposed scoring algorithm to predict comorbidity. The LDSC-DDN had a higher density value of 19.97% (18 165 edges across 427 nodes) than the signed DDN

Table 1. The top 20 diseases having direct comorbidity with coronary atherosclerosis

| Phenotypes (PheCode) | Scores | RR [95% CI] | ϕ -corr |
|--|--------|-------------------------|--------------|
| Myocardial infarction (411.2) | 1.000 | 13.052 [13.050, 13.054] | 0.5564 |
| Other chronic ischemic heart disease, unspecified (411.8) | 0.993 | 11.636 [11.634, 11.637] | 0.5521 |
| Angina pectoris (411.3) | 0.929 | 11.510 [11.508, 11.511] | 0.5573 |
| Unstable angina (411.1) | 0.855 | 13.261 [13.256, 13.266] | 0.3512 |
| Cardiac congenital anomalies (747.1) | 0.840 | 4.115 [4.000, 4.234] | 0.0185 |
| Cardiac and circulatory congenital anomalies (747)* | 0.831 | 0.918 [0.344, 2.445] | -0.0002 |
| Congenital anomalies of great vessels (747.13) | 0.827 | 8.758 [8.746, 8.770] | 0.1439 |
| Other cerebral degenerations (331) | 0.811 | 2.432 [2.348, 2.519] | 0.0100 |
| Primary open angle glaucoma (365.11) | 0.790 | 1.620 [1.601, 1.639] | 0.0092 |
| Open-angle glaucoma (365) | 0.788 | 1.614 [1.595, 1.632] | 0.0091 |
| Delirium dementia and amnestic and other cognitive disorders (290) | 0.777 | 3.311 [3.261, 3.361] | 0.0211 |
| Altered mental status (292.4) | 0.756 | 2.880 [2.871, 2.889] | 0.0400 |
| Dizziness and giddiness (386.9) | 0.738 | 3.114 [3.110, 3.119] | 0.0624 |
| Heart valve replaced (395.6) | 0.734 | 9.291 [9.276, 9.306] | 0.1371 |
| Aortic valve disease (394.3) | 0.729 | 7.573 [7.557, 7.590] | 0.1042 |
| Glaucoma (365) | 0.724 | 1.882 [1.876, 1.888] | 0.0234 |
| Cerebral ischemia (433.3) | 0.710 | 3.670 [3.650, 3.690] | 0.0385 |
| Occlusion and stenosis of precerebral arteries (433.1) | 0.706 | 6.338 [6.316, 6.361] | 0.0718 |
| Vertiginous syndromes and other disorders of vestibular system (386) | 0.684 | 2.008 [1.681, 2.400] | 0.0034 |
| Neurological disorders (292) | 0.580 | 2.782 [2.741, 2.825] | 0.0177 |

Note: An asterisk denotes a disease pair that were predicted as having direct comorbidity from DDN-driven comorbidity, but inverse comorbidity on the basis of EHR-driven comorbidity (RR, relative risk; 95% CI, 95% confidence interval; corr, correlation).

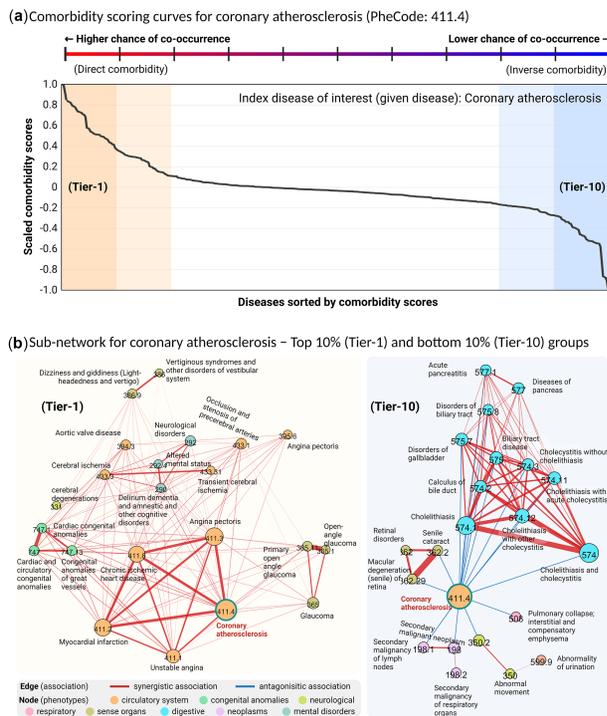


Fig. 6. Comorbidity scoring curves and associations for coronary atherosclerosis. (a) Comorbidity scoring curves for coronary atherosclerosis (PheCode: 411.4). The solid line indicates DDN-driven comorbidity scores (f) of the other 426 diseases in the dataset, sorted in descending order with respect to score on the x-axis. The y-axis represents comorbidity scores, rescaled to values between -1 and 1 . (b) Sub-networks of diseases having high comorbidity (Tier-1) and low comorbidity (Tier-10) with coronary atherosclerosis

(density of 10.14% as shown in Fig. 3c). We performed comorbidity prediction tasks and calculated AUCs with the same experimental settings as the signed network. Empirically, the dense network can provide more accurate inferences than sparse network. However, comorbidity prediction for 98 diseases could not be conducted because they were disconnected from other diseases in the LDSC

DDN. The LDSC results cannot be obtained when applied to datasets with low statistical power of GWAS (due to small sample size or rare trait) for estimating heritability and genetic correlation. Most connections in the LDSC-DDN were obtained from diseases with high statistical power. The proposed signed DDN had advantage of discovering inter-disease interactions by leveraging significantly associated SNPs, even though the number of SNPs used for constructing network was smaller. The detailed results are provided in [Supplementary Text S3](#).

4 Discussion and conclusion

We proposed a novel signed DDN based on biobank-scale PheWAS summary statistics that considers the direction of effect of shared SNPs, and further presented the utility of this DDN in prioritizing diseases according to comorbidity. To design this DDN, we measured the overall direction of effect of SNPs shared between pairs of diseases, and further categorized disease–disease associations as synergistic or antagonistic depending on whether that overall direction is consistent or opposite between diseases. Our results demonstrated that considering the interaction and direction of shared components is more helpful in predicting comorbidity and ranking comorbid diseases than considering only the quantity of shared components.

We also developed a novel label propagation algorithm for signed networks and applied it to show that the signed DDN not only models disease relationships at the population level, but also is applicable to comorbidity prediction in the context of personalized medicine. The comorbidity scoring algorithm was designed such that when the initial label on the index disease encounters a ‘disease connected by a synergistic association’, the score is propagated with a positive value (increasing the likelihood of co-occurrence), whereas when it encounters a ‘disease connected by an antagonistic association’, the score is propagated with a negative value (reducing the likelihood of co-occurrence).

One advantage of the proposed method is that the signed DDN can aid in the interpretation and comprehension of the intricate associations among diseases, and moreover can be easily updated and reinforced when new PheWAS results are obtained for a phenotype not yet included in the network. A second advantage is its ability to categorize diseases as having direct and inverse comorbidity based on the defined synergistic/antagonistic associations. It is very

important to know both which conditions have higher chance of co-occurrence (direct comorbidity) and which have lower chance of co-occurrence (inverse comorbidity) for a given disease. Such categorized relationships enable clinicians to construct comorbidity scenarios in advance, and can aid patient treatment planning. Our study also used co-occurrence and disease prevalence information obtained from UK Biobank hospitalization episodes to confirm and validate the DDN-driven predictions of direct/inverse comorbidity. Although omnidirectional validation has not been performed, our findings support that a DDN based on the relations of diseases and genetic components provides sufficient information for the prioritization and categorization of disease comorbidity.

Some aspects of this study remain to be investigated in future work. One of the primary limitations is that we utilized data from a single cohort of the UK Biobank. Since the UK Biobank mainly includes healthy participants, the constructed network might not capture disease prevalence in populations of different demographic characteristics. We also limited our analyses to those phenotypes occurring in both sexes. Some significant genetic components might vary by sex, and thus it is necessary to construct sex-specific signed DDNs and determine sex-specific comorbidity. Although our stratification of comorbidity groups is a product of a specific population, the approach can be extended for use in precision medicine to screen individuals at comorbidity risks.

Acknowledgements

We acknowledge all the participants of the UK Biobank. The use of the UK Biobank resource was approved under Application Number 33002. We thank Lee Lab (Prof. Seunggeun Lee) for providing UK Biobank PheWAS summary data for these analyses. We thank the Associate Editor and the referee for constructive comments, which have improved the article. All figures were created or edited using BioRender.com (accessed on October 30, 2021).

Funding

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No. 2022R1A2C2009998]; the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea [HU21C0111]; and by the National Institute of General Medical Sciences (NIGMS) [R01 GM138597].

Conflict of Interest: none declared.

Data availability

The UK Biobank summary data are publicly available at <https://www.lee-lab.org/resources>. The source codes and data are available on <https://github.com/dokyoonkimlab/DiseaseInteractiveNetwork>, the implementation of inter-disease interactivity network is available on <https://hdpm.biomedinfo-lab.com/ddn/signedDDN>.

References

Auton, A. *et al.*; 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Barabási, A.-L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Buddeke, J. *et al.* (2019) Comorbidity in patients with cardiovascular disease in primary care: a cohort study with routine healthcare data. *Br. J. Gen. Pract.*, **69**, e398–e406.

Bulik-Sullivan, B. *et al.*; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3. (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236–1241.

Catalá-López, F. *et al.* (2014) Inverse and direct cancer comorbidity in people with central nervous system disorders: a meta-analysis of cancer incidence in 577,013 participants of 50 observational studies. *Psychother. Psychosom.*, **83**, 89–105.

Cho, S.I. *et al.* (2021) Impact of comorbidity burden on mortality in patients with COVID-19 using the Korean health insurance database. *Sci. Rep.*, **11**, 1–9.

Chong, Y. *et al.* (2020) Graph-based semi-supervised learning: a review. *Neurocomputing*, **408**, 216–230.

Darabos, C. *et al.* (2015) A bipartite network approach to inferring interactions between environmental exposures and human diseases. *Pac. Symp. Biocomput.*, **2015**, 171–182.

Di Angelantonio, E. *et al.*; Emerging Risk Factors Collaboration. (2009) Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*, **302**, 1993–2000.

Dong, G. *et al.* (2021) A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank. *Genome Med.*, **13**, 110.

Fan, Q. *et al.* (2017) HDL-cholesterol levels and risk of age-related macular degeneration: a multiethnic genetic study using Mendelian randomization. *Int. J. Epidemiol.*, **46**, 1891–1902.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Fischer, M. *et al.* (2005) Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. *Circulation*, **111**, 855–862.

Gallier, J. (2016) Spectral theory of unsigned and signed graphs. Applications to graph clustering: a survey. arXiv:1601.04692.

Goh, K.-I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.

Hidalgo, C.A. *et al.* (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.*, **5**, e1000353.

Ibáñez, K. *et al.* (2014) Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. *PLoS Genet.*, **10**, e1004173.

Jørgensen, T.L. *et al.* (2012) Comorbidity in elderly cancer patients in relation to overall and cancer-specific mortality. *Br. J. Cancer*, **106**, 1353–1360.

Klimek, P. *et al.* (2015) Quantification of diabetes comorbidity risks across life using nation-wide big claims data. *PLoS Comput. Biol.*, **11**, e1004125.

Lee, D.-G. *et al.* (2020) Dementia key gene identification with multi-layered SNP-gene-disease network. *Bioinformatics*, **36**, i831–i839.

Lee, D.S. *et al.* (2008) The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. USA*, **105**, 9880–9885.

Liu, C.C. *et al.* (2014) DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Res.*, **42**, W137–W146.

Marenberg, M.E. *et al.* (1994) Genetic susceptibility to death from coronary heart disease in a study of twins. *N. Engl. J. Med.*, **330**, 1041–1046.

McPherson, R. and Tybjaerg-Hansen, A. (2016) Genetics of coronary artery disease. *Circ. Res.*, **118**, 564–578.

Nam, Y. *et al.* (2019a) Disease gene identification based on generic and disease-specific genome networks. *Bioinformatics*, **35**, 1923–1930.

Nam, Y. *et al.* (2019b) The translational network for metabolic disease—from protein interaction to disease co-occurrence. *BMC Bioinformatics*, **20**, 12.

Nam, Y. *et al.* (2022) netCRS: network-based comorbidity risk score for prediction of myocardial infarction using biobank-scaled PheWAS data. *Pac. Symp. Biocomput.*, **27**, 325–336.

Nashiry, A. *et al.* (2021) Bioinformatics and system biology approach to identify the influences of COVID-19 on cardiovascular and hypertensive comorbidities. *Brief. Bioinform.*, **22**, 1387–1401.

Richardson, S. *et al.*; the Northwell COVID-19 Research Consortium. (2020) Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *JAMA*, **323**, 2052–2059.

Roitmann, E. *et al.* (2014) Patient stratification and identification of adverse event correlations in the space of 1190 drug related adverse events. *Front. Physiol.*, **5**, 332.

Rubio-Perez, C. *et al.* (2017) Genetic and functional characterization of disease associations explains comorbidity. *Sci. Rep.*, **7**, 1–14.

Sánchez-Valle, J. *et al.* (2020) Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. *Nat. Commun.*, **11**, 2854.

Schisterman, E.F. *et al.* (2005) Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, **16**, 73–81.

Sharrett, A.R. *et al.*; Atherosclerosis Risk in Communities Study Group. (2001) Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins A-I and B, and HDL density

- subfractions: the atherosclerosis risk in communities (ARIC) study. *Circulation*, **104**, 1108–1113.
- Subramanya,A. and Talukdar,P.P. (2014) Graph-based semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.*, **8**, 1–125.
- Tabarés-Seisdedos,R. and Baudot,A. (2016) Direct and inverse comorbidities between complex disorders. *Front. Physiol.*, **7**, 117.
- Tabarés-Seisdedos,R. and Valderas,J.M. (2013) Inverse comorbidity: the power of paradox in the advancement of science. *J. Comorb.*, **3**, 1–3.
- Tarantino,G. et al. (2018) Prediction of carotid intima-media thickness in obese patients with low prevalence of comorbidities by serum copper bioavailability. *J. Gastroenterol. Hepatol.*, **33**, 1511–1517.
- Valderas,J.M. et al. (2009) Defining comorbidity: implications for understanding health and health services. *Ann. Fam. Med.*, **7**, 357–363.
- Verma,A. et al.; DiscovEHR Collaboration. (2019) Human-disease phenotype map derived from PheWAS across 38,682 individuals. *Am. J. Hum. Genet.*, **104**, 55–64.
- von Mutius,E. and Smits,H.H. (2020) Primary prevention of asthma: from risk and protective factors to targeted strategies for prevention. *Lancet*, **396**, 854–866.
- Wei,W.-Q. et al. (2017) Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One*, **12**, e0175508.
- Wu,P. et al. (2019) Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med. Inform.*, **7**, e14325.
- Yule,G.U. (1919) *An Introduction to the Theory of Statistics*. C. Griffin, Limited, London, UK.
- Zheng,J. et al. (2018) PhenoSpD: an integrated toolkit for phenotypic correlation estimation and multiple testing correction using GWAS summary statistics. *Gigascience*, **7**, giy090.
- Zheng,J. et al.; Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, **33**, 272–279.
- Zhou,W. et al. (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.*, **50**, 1335–1341.
- Zhou,X. et al. (2014) Human symptoms-disease network. *Nat. Commun.*, **5**, 4212.
- Zhu,Z. et al. (2018) Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.*, **9**, 12.