HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

Pac Symp Biocomput. 2022 ; 27: 325–336.

netCRS: Network-based comorbidity risk score for prediction of myocardial infarction using biobank-scaled PheWAS data

Yonghyun Nam[1,§], **Sang-Hyuk Jung**[1,2,§], **Anurag Verma**[3], **Vivek Sriram**[1], **Hong-Hee Won**[2], **Jae-Seung Yun**[1,4], **Regeneron Genetics Center**, **Dokyoon Kim**[1,5,*]

[1]Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[2]Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul 06351, Republic of Korea

[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[4]Division 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

[5]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

The polygenic risk score (PRS) can help to identify individuals' genetic susceptibility for various diseases by combining patient genetic profiles and identified single-nucleotide polymorphisms (SNPs) from genome-wide association studies. Although multiple diseases will usually afflict patients at once or in succession, conventional PRSs fail to consider genetic relationships across multiple diseases. Even multi-trait PRSs, which take into account genetic effects for more than one disease at a time, fail to consider a sufficient number of phenotypes to accurately reflect the state of disease comorbidity in a patient, or are biased in terms of the traits that are selected. Thus, we developed novel network-based comorbidity risk scores to quantify associations among multiple phenotypes from phenome-wide association studies (PheWAS). We first constructed a disease-SNP heterogeneous multi-layered network (DS-Net), which consists of a disease network (disease-layer) and SNP network (SNP-layer). The disease-layer describes the population-level interactome from PheWAS data. The SNP-layer was constructed according to linkage disequilibrium. Both layers were attached to transform the information from a population-level interactome to individual-level inferences. Then, graph-based semi-supervised learning was applied to predict possible comorbidity scores on disease-layer for each subject. The SNP-layer serves as receiving individual genotyping data in the scoring process, and the disease-layer serves as the propagated output for an individual's multiple disease comorbidity scores. The possible comorbidity scores were combined by logistic regression, and it is denoted as netCRS. The DS-Net was constructed from UK Biobank PheWAS data, and the individual genetic profiles were

* Dokyoon.kim@pennmedicine.upenn.edu .
§Co-first author

collected from the Penn Medicine Biobank. As a proof-of-concept study, myocardial infarction (MI) was selected to compare netCRS with the PRS with pruning and thresholding (PRS-PT). The combined model (netCRS + PRS-PT + covariates) achieved an AUC improvement of 6.26% compared to the (PRS-PT + covariates) model. In terms of risk stratification, the combined model was able to capture the risk of MI up to approximately eight-fold higher than that of the low-risk group. The netCRS and PRS-PT complement each other in predicting high-risk groups of patients with MI. We expect that using these risk prediction models will allow for the development of prevention strategies and reduction of MI morbidity and mortality.

Keywords

Comorbidity; polygenic risk scores; graph-based semi-supervised learning; multi-layered network

1. Introduction

The prediction of an individual's disease risk is an essential part of precision medicine and will be required to improve public healthcare and understand risk of developing a disease across different populations. One of the most popular methods of disease risk prediction is the polygenic risk score (PRS), which estimates a patient's genetic risk for a chosen trait or disease by combining individual genetic profiles with many single-nucleotide polymorphisms (SNPs) identified through genome-wide association studies (GWAS).[1,2] Many studies have calculated PRSs for various common diseases, including cardiovascular disease, hypertension, and neurological disorders, and they suggest that the PRS might be a helpful tool for identifying and categorizing high-genetic risk individuals for those diseases.[3-6] Nevertheless, a major weakness of the conventional PRS is its focus on a single trait for the estimation of genetic risk scores – when predicting the risk scores of an index disease of interest, PRS is calculated according solely to the relevant phenotype. In most cases, however, multiple diseases will usually afflict a patient at once or in succession. These disease complications and comorbidities, referring to the presence of one or more additional medical conditions given a primary disease, suggest that effective disease prediction will require us to consider multiple phenotypes concurrently.[7] In order to estimate the disease risk considering the associations among multiple diseases, several studies had attempted to perform the association analysis for PRSs with multiple diseases through subsequent analysis[8,9] or to combine PRSs for multiple traits.[10] In these previous studies, a key step involves the determination of which diseases related to the index disease are selected for estimation of the combined risk score. However, these methods are limited as selection bias is introduced when knowledge reveled in clinical practice is used to identify diseases highly related to the target phenotype. Even multi-trait PRSs, which take into account genetic effects for more than one disease at a time, fail to consider a sufficient number of phenotypes to accurately reflect the state of disease comorbidity in a patient, or are biased in terms of the traits that are selected.

One effective way to comprehensively explore the genetic associations among multiple diseases is to consider a network representation, such as the disease-disease network (DDN). Given a set of diseases, the DDN represents diseases as nodes, and disease-

disease associations as edges. DDNs can explore potential comorbidity relationships among phenotypes based on shared genetic components. Different genetic components will yield different types of networks, such as gene[11], protein[12], pathway[13], and SNP-based DDN.[14] In this study, the SNP-based DDN is used to incorporate the conventional PRS approach, where edges represent the number of shared SNPs between diseases according to results from a phenome-wide association study (PheWAS). The SNP-based DDN using PheWAS results is depicted in the center panel of Figure 1. Considering D2 as an index disease of interest (marked in red), we can see that it is directly connected with four diseases (D1, D3, D4, and D6). Three diseases (D5, D7, and D8) share no edges with D2. Directly connected diseases share at least one common SNP with D2. Indirectly connected diseases share no genetic associations with D2, but they are connected through the other nodes – for instance, D2 and D7 are connected in through the sequence of diseases with D2~D6~D7. Overall population-level relationships between diseases can be observed through the underlying structure of the DDN, regardless of whether or not a pair of diseases share genetic components. In developing risk prediction models which consider the relationships across a multitude of diseases, a DDN can provide intuitive, unbiased evidence about the selection of related diseases as well as the strength of associations between diseases. However, although the population-level interactome between phenotypes can be observed through a DDN, it is not easy to apply these disease-disease associations in a patient-specific manner. Indeed, it is difficult to obtain information pertinent to the individual because the nodes and edges in DDN are aggregated and summarized from PheWAS data.

To circumvent this challenge, we propose a novel framework of network-based individual comorbidity risk scores (netCRS) to predict individual-level disease comorbidity risk through population-level interactome networks. The goals of netCRS are as follows: (a) To improve the prediction ability of PRS, we present a novel risk score that estimates multiple disease comorbidities according to their shared genetic components. The netCRS estimates the combined comorbidity scores for multiple phenotypes in the SNP-based DDN when provided with an individual genetic profile. In PRS, marginal effect size estimates of SNPs obtained from a GWAS are used as weights for weighted sum scores of risk alleles carried by an individual for a single trait. On the other hand, in netCRS, disease-specific effect size estimates of SNPs from PheWAS are used as edge weights of the network for multiple traits. (b) To obtain individual-level inference from population-level interactome, we construct a novel disease-SNP heterogeneous multi-layered network using EHR-linked biobank-scale PheWAS summary statistics. Using this multi-layered network, we introduce a scoring method to infer individual information from population-level networks through layer-wise label propagation.

Figure 1 describes the overall conceptual framework of netCRS. The center panel depicts a disease-SNP heterogeneous multi-layered network (denoted as DS-Net). The DS-Net is a multi-layered graph, consisting of a SNP-SNP correlation network (SNP-layer), disease-disease network (disease-layer) and SNP-disease associations (coupling graphs). Briefly, the SNP-layer (colored solid circles/lines) is constructed according to a linkage disequilibrium matrix, and the disease-layer (colorless solid circles/lines) is constructed according to the shared genetic components between phenotypes. The coupling graphs for inter-layers (colored dashed lines) between the SNP- and disease-layer are derived using

disease-SNP associations obtained from PheWAS summary statistics. Given the DS-Net and index disease of interest, we first predict individual comorbidity scores using graph-based semi-supervised learning (SSL). Graph-based SSL predicts scores on the disease-layer by propagating label information when the individual genetic profile is labeled on the SNP layer. In the left panel of Figure 1, individual genotype data is used to provide query or seed label information to the SNP-layer for the scoring algorithm. Each patient's genetic data are initially labeled on the SNP-layer, and then the label information is propagated through the multi-layered network. Predicted risk scores are obtained for each disease node (blue bar). Each bar depicts a possible comorbidity score for each disease that an individual patient can have. The predicted comorbidity scores are subsequently aggregated into combined comorbidity scores using a meta-classifier (the right panel of Figure 1). Here, we use logistic regression for our meta-learner, and the combined comorbidity score is denoted as netCRS( ), where the parentheses specify the index disease of interest. More details of the proposed methods are explained in the following sections.

2. netCRS: Network-based individual Comorbidity Risk Scoring

2.1. Disease-SNP Heterogeneous Network using UK Biobank summary statistics

We constructed the reference network using UK BioBank (UKBB) PheWAS summary statistics. The DS-Net is a multi-layered weighted graph, $\boldsymbol{G = (V, W, S)}$, where $\boldsymbol{V}$ represents the set of nodes, $\boldsymbol{W}$ represents the set of edges, and $\boldsymbol{S}$ represents the set of layers. The multi-layered network $\boldsymbol{G}$ is decomposed into two distinct single graphs with corresponding layers $S = \{S_{\text{Disease}}, S_{\text{SNP}}\}$. The similarity matrix $\boldsymbol{W}$ for multi-layered network can be expressed in block-wise matrix as follows:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{Disease} & \boldsymbol{C} \\ \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{W}_{SNP} \end{bmatrix} \tag{1}$$

The block diagonal matrix ($\boldsymbol{W}_{\text{Disease}}$ and $\boldsymbol{W}_{\text{SNP}}$) represents a similarity matrix for each single network of the disease-layer and SNP-layer respectively, and the block-off diagonal matrix $\boldsymbol{C}$ represents the coupling graphs for the connections between inter-layers.

2.1.1. Disease-Layer (Disease-Disease network)—The disease-layer $\boldsymbol{G}_{\text{Disease}} = (\boldsymbol{V}_{\text{Disease}}, \boldsymbol{W}_{\text{Disease}})$ is a sub-network of the DS-Net $\boldsymbol{G}$, where the nodes $\boldsymbol{V}_{\text{Disease}}$ denotes the set of dseases, and $\boldsymbol{W}_{\text{Disease}}$ de notes the similarity between the sequences of SNPs that pairs of diseases commonly share. The disease-layer is constructed according to shared genetic components, with the hypothesis that two different phenotypes are associated if they share significant SNPs from the PheWAS summary results. Given m diseases and k SNPs, we first generate m disease-SNP as sociation vectors from each PheWAS result. Each disease vector $\boldsymbol{v}$ is represented as a k-dimensional SNP vector with binary attributes, each of which stands for statistically significant ('1') or not ('0') for the association with a specific SNP that has passed the p-value thresholds in the PheWAS results.[14] Then, similarity between pairs of diseases is measured by cosine similarity w_{ij} for two diseases v_i and v_j.

$$w_{ij}^{\text{Disease}} = \frac{\boldsymbol{v_i} \cdot \boldsymbol{v_j}}{\|\boldsymbol{v_i}\| \cdot \|\boldsymbol{v_j}\|} \quad (2)$$

2.1.2. SNP-layers (SNP-SNP correlation network)—SNP-layer $\boldsymbol{G}_{\text{SNP}} = (\boldsymbol{V}_{\text{SNP}}, \boldsymbol{W}_{SNP})$ is a sub-network of the disease-SNP heterogeneous network $\boldsymbol{G}$ when $\boldsymbol{S} = \{S_{\text{SNP}}\}$. The node $\boldsymbol{V}_{\text{SNP}}$ denotes the representative SNPs after genetic pre-processing, and $\boldsymbol{W}_{\text{SNP}}$ denotes the pairwise genetic correlations between distinct SNPs. We generate the pairwise linkage-disequilibrium (LD) matrices of genotype correlation between nearby SNPs using quality-controlled genotyped data of UKBB samples. The r^2 between pairs of SNPs is obtained using PLINK 1.90 with LD calculation (--r2, --1d-window 10 SNPs, --ld-window-kb 1000kb, and --ld-window-r2 0.0). The similarity matrix $\boldsymbol{W}_{\text{SNP}}$ is composed of correlation values ranging from 0 to 1.

2.1.3. Coupling graphs (SNP-Disease associations)—The coupling graphs $\boldsymbol{C} = \{c_{ik} | i \in \boldsymbol{V}_{Disease}, k \in \boldsymbol{V}_{SNP}\}$ imply connections between diseases and SNPs across different layers of the network. Coupling edges are derived from the disease-SNP association vectors (described in section 2.1.1). Edge weights take value of z-scores, equivalent to the beta-coefficients (β_{ik}) divided by standard errors (SE_{ik}) from the significant association between phenotype i and SNP k from PheWAS results. These weights are rescaled to lie within a range of 0 to 1.

Combining the disease-layer, SNP-layer, and coupling graphs yields the proposed DS-Net. The constructed network can provide insights into the population-level interactome between diseases and SNPs.

2.2. Individual comorbidity risk scoring algorithms

Given an index disease of interest, we can predict individuals' disease comorbidity risk scores using the DS-Net. Since the network describes a biobank-scale population-level interactome, we take individual genetic information from another biobank to calculate risk scores for individual patients. In this analysis, the summary-level data from UKBB were used for the network construction, and the individual-level genetic data were collected from the Penn Medicine BioBank (PMBB). More details are explained in the Section 3.

Let us define disease comorbidity risk scoring $\boldsymbol{f}: \mathbf{V} \rightarrow \mathbb{R}$ as a function that quantifies the degree of commitment of the diseases associated with SNPs on the network. To implement this scoring function in a DS-Net, we employ graph-based SSL with transductive learning settings.[15] As shown in Figure 1, individual genotypes are used for initial label information in the DS-Net. We set the genotype CC (homozygous non-reference) as 0, genotype CT (heterozygous) as 0.5, and genotype TT (homozygous reference) as 1 for initial labels of label propagation. Once the labels for the SNP-layer are provided, graph-based SSL propagates the label information through all edges in the heterogeneous multi-layered network simultaneously. Since we are interested only in the comorbidity risk of multiple diseases, the propagated disease scores $\boldsymbol{f}_{\text{Disease}}$ on the disease-layer $\boldsymbol{V}_{\text{Disease}}$ are used as

the predicted comorbidity feature vectors. To aggregate these scores, we employ logistic regression as the meta-classifier.

The following section describes the formulation of the proposed network-based comorbidity scoring algorithm. Assume that we have genotype data for m individuals and that we know the diagnosis outcomes of the index disease. Then, i-th patient's genotype data $\boldsymbol{m}_i$ has k-dimensional SNP vectors with values of 0, 0.5, and 1 as described above. The outcomes of the index disease for all patients $\boldsymbol{z}$ is an m-dimensional vector with value '1' if the patient has been diagnosed with the index disease or '0' otherwise. To apply the individual data to graph-based SSL, we set the initial label set of vector $\boldsymbol{y}$ and predicted scores $\boldsymbol{f}$. The initialization and learning process is performed iteratively patient-by-patient. Let $\boldsymbol{y} = (y_1, \dots, y_n, y_{n+1}, \dots, y_{n+k})^{\mathrm{T}} = (\boldsymbol{y}_{\mathrm{Disease}}, \boldsymbol{y}_{\mathrm{SNP}})^{\mathrm{T}}$ denote the set of initial labels and $\boldsymbol{f} = (f_1, \dots, f_n, f_{n+1}, \dots, f_{n+k})^{\mathrm{T}} = (\boldsymbol{f}_{\mathrm{Disease}}, \boldsymbol{f}_{\mathrm{SNP}})^{\mathrm{T}}$ den ote the set of predicted scores, where n is the total number of diseases and k is the total number of SNPs in the network. In the problem setting of disease comorbidity scores, we set the $\boldsymbol{y}_{\mathrm{Disease}}$ to the zero vector and $\boldsymbol{y}_{\mathrm{SNP}}$ to $\boldsymbol{m}_i$. The label information is propagated to all connected nodes along with edges in $\boldsymbol{W}_{\mathrm{SNP}}$, $\mathbf{C}$, and $\mathbf{W}_{\mathrm{Disease}}$ on graph $\mathbf{G}$. Graph-based SSL provides the real-valued scores $\boldsymbol{f}$ with two assumptions: (a) smoothness function (predicted scores f_i and f_j should not be different if two nodes v_i and v_j are adjacent), (b) loss function (predicted scores f_i should be close to the given label of y_i). We can obtain predicted score $\boldsymbol{f}$ by minimizing the following quadratic function:

$$\min_{f} \ (\boldsymbol{f} - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f} - \boldsymbol{y}) + \mu \boldsymbol{f}^{\mathbf{T}} \boldsymbol{L} \boldsymbol{f} \tag{3}$$

where $\boldsymbol{L}$ is the graph Laplacian defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, $\boldsymbol{D} = \mathrm{diag}(d_i)$ is diagonal degree matrix, $d_i = \Sigma_j w_{ij}$, and μ is user-specific parameter that provides a trade-off between the loss function (first term of Eq. (3)) and smoothness function (second term of Eq. (3)). The closed form of solution $\boldsymbol{f}$ becomes

$$\boldsymbol{f} = (\boldsymbol{I} + \mu \boldsymbol{L})^{-1} \boldsymbol{y} \tag{4}$$

The predicted scores f on Eq. (4) can be re-expressed in a block-wise representation by using Eq. (1).[12]

$$\begin{bmatrix} \boldsymbol{y}_{\mathrm{Disease}} \\ \boldsymbol{y}_{\mathrm{SNP}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} + \mu(\boldsymbol{D}_{\mathrm{Disease}} - \boldsymbol{W}_{\mathrm{Disease}}) & -\mu \boldsymbol{C} \\ -\mu \boldsymbol{C}^{\mathrm{T}} & \boldsymbol{I} + \mu(\boldsymbol{D}_{\mathrm{SNP}} - \boldsymbol{W}_{\mathrm{SNP}}) \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_{\mathrm{Disease}} \\ \boldsymbol{f}_{\mathrm{SNP}} \end{bmatrix} \tag{5}$$

Since the nodes in the SNP-layer are all labeled and nodes in the disease-layer are all unlabeled, Eq. (5) is simplified by substituting $\boldsymbol{f}_{\mathrm{SNP}}$ as $\boldsymbol{y}_{\mathrm{SNP}}$ and $\boldsymbol{y}_{\mathrm{Disease}}$ as $\mathbf{0}$. The predicted scores on the disease-layer are thus obtained as

$$\boldsymbol{f}_{\mathrm{Disease}} = \mu\{\boldsymbol{I} + \mu(\boldsymbol{D}_{\mathrm{Disease}} - \boldsymbol{W}_{\mathrm{Disease}})\}^{-1} \boldsymbol{C} \cdot \boldsymbol{y}_{\mathrm{SNP}} \tag{6}$$

This process is iteratively repeated for each individual patient, and $\boldsymbol{F}^* = \{\boldsymbol{f}^{(1)}_{\text{Disease}}, \ldots, \boldsymbol{f}^{(m)}_{\text{Disease}}\}$ represents the m-dimensional comorbidity score vector. To aggregate these vectors, we employ logistic regression as a meta-classifier with $\mathbf{z} \sim \boldsymbol{\beta}^{\text{T}}\boldsymbol{f}_{\text{Disease}} + \boldsymbol{\epsilon}$. We can then obtain the combined possible comorbidity risk scores as $\textbf{netCRS}(\,) = \widehat{\boldsymbol{\beta}}^{\text{T}}\boldsymbol{F}^*$ for the individual. A step-by-step process for scoring is summarized with pseudo-code in Supplementary Figure 1.

3. Results

In this study, we selected myocardial infarction (PheCode: 411.2) as the index disease of interest. It is commonly known as a heart attack and occurs when blood flow reduces or stops to a part of the heart. Myocardial infarction (MI) is the main undesirable outcome of coronary artery disease. Coronary artery disease, often caused by coronary atherosclerosis, is a common chronic condition characterized by a substantial and complex polygenic contribution to disease risk, with a heritability between 40% and 60%. We describe a Mi-specific DS-Net and present comorbidity scores of MI for the individual, **netCRS** (*myocardial infarction, MI*).

3.1. Experimental Setting

3.1.1. Data for model development and validation set—To build the Mi-specific DS-Net and calculate netCRS(MI), a total of 1,403 PheCode-based UK biobank PheWAS summary statistics were obtained from https://www.leelabsg.org/resources.[16] To construct the myocardial infarction-specific DS-Net, 135 diseases were selected with the following criteria: (a) The diseases were included in the disease-layer if phenotypes had a minimum number of cases larger than 1000, and (b) the diseases were included if phenotypes had at least one shared SNP with myocardial infarction (directly connected with MI). The selected disease categories and disease-layers are described in Figure 2. In the SNP-layer, 39,365 SNPs were selected with genome-wide significance p-value threshold $\leq 1 \times 10^{-4}$. Linkage disequilibrium (LD) pruning was performed with thresholds (window size: 50, step size: 5, and r^2 threshold: 0.5). A list of components in the DS-Net is described in Supplementary Table 1.

Individual genotype data were collected from the PMBB. The PMBB is an institutional research program that recruits patient-participants throughout the University of Pennsylvania Health System by enrolling at the time of outpatient visits ore more recently, through electronic consenting. Approximately 45,000 of these participants already have genotype data available along with electronic health records (EHR). ICD-9 and ICD-10 codes were aggregated to PheCodes by referring to the PheCode Map 1.2 version.[17-19] 4,972 individuals of European ancestry were included for this study, all of whom underwent genotyping and had available electronic health record data (Table 1). The detailed genotype QC we performed refers to the previous study [20]. According to the accumulated medical history at the time of participation, individuals were considered cases for MI if they had at least 2 instances of the PheCode on unique dates, controls if they had no instance of the PheCode, and ‘other/missing’ if they had one instance or a related PheCode. Table 1 describes the list of data and sources for model development and validation cohort.

3.1.2. Experimental Setting—To evaluate the prediction performance of netCRS using PMBB genotype data, we compared proposed method to PRS with pruning and thresholding (PRS-PT), calculated using PRSice-2[21]. Area under the receiver operating characteristic curve (AUC) was used as performance measure. The model parameters were searched over the following ranges for the respective models. In netCRS(MI), we performed a hyper-parameter search of μ for Eq. (4) of graph-based SSL over $\mu = \{0.01, 0.1, 1, 10, 100\}$. The PRS-PT was generated from the sum of the risk alleles weighted by their effect sizes based on GWAS summary statistics from Coronary Artery Disease Genome-wide Replication and Meta-analysis plus the Coronary Artery Disease Genetics (CARDIOGRAMplus C4D consortium).[22] The parameters were selected from a range of p-value thresholds $\{5 \times 10^{-8}, 1 \times 10^{-6}, 0.0001, 0.001, 0.01, 0.05\}$ and LD-based clumping r^2 (0.1 to 0.9) within 1,000 kb. The generated netCRS(MI) and PRS-PT(MI) were compared between MI cases and healthy controls with the logistic regression model, respectively. For both models, the best performance was selected by searching over the respective model-parameter space. The best model of PRS-PT(MI) was determined based on the optimal threshold with the largest Nagelkerke's R^2 value (in Supplementary Table 1).

3.1.3. Risk predictions of myocardial infarction with netCRS—Table 2 shows the performance comparison of the best PRS-PT(MI) and netCRS(MI) in terms of overall AUC for MI cases and healthy controls. In the results, we included the prediction performance of singleton risk model (netCRS and PRS-PT) and models with covariates of sex and age. We also included the additive models of (PRS-PT + netCRS) with and without covariates. The netCRS with $\mu = \{0.1\}$ achieved best predictive performance a cross both singleton and additive models. When netCRS was used along with the conventional PRS model, the combined model [6] (netCRS + PRS-PT + covariates) achieved an AUC improvement of 28.29%(=(0.7417 − 0.5827)/0.5827) compared to the PRS-PT alone model [1] in MI case prediction. Also, the combined model [6] improved the performance up to 0.7417 of AUC (lifted from 0.6979), comparing to the individual PRS-PT model [4] (AUC improvement of 6.26%). Models with superscript of asterisk were used in further association analysis to validate netCRS and its effectiveness (model [2], [5], and [6])

3.1.4. Association analysis of netCRS and PRS—To investigate the effectiveness of the association between both risk scoring models and covariates with age and sex, we assessed multiplicative interactions between netCRS and each of the stratification variables. We stratified participants based on quartiles of netCRS; low risk (0th-25th), intermediate risk (26th-50th), high risk (51st-75th), and very high risk (76th-100th). Compared with the low-netCRS risk group, the higher netCRS risk group had higher odds ratios in the validation cohort. In stepwise multivariate models (model [5] and [6]), the models with covariates and/or PRS-PT remained significantly (Table 3). Participants in the very high-netCRS risk group for MI had approximately four-fold increased risk of MI occurrence relative to those with the corresponding low-genetic risk group (shown in Table 3). In addition, we investigated the benefit of using netCRS and PRS together in screening high-risk groups for MI. Table 4 demonstrates that combinations of MI-PRS and netCRS were able to capture the risk of MI up to approximately eight-fold higher than the low-risk group.

Supplementary Table 3 provides demographics of participants according to netCRS risk groups.

4. Conclusion

In this study, we developed and proposed netCRS, a network-based disease comorbidity risk scoring algorithm based upon biobank-scale PheWAS summary statistics. To improve the prediction ability of PRS, we introduced a novel combined comorbidity risk scores using a multi-layered network. Most current biological networks suggest only associative information between biological components according to aggregated population-level data [23]. Although these population-level networks provide insights regarding the interaction of components, it is not easy to obtain individual inference from them.

To solve this problem, we proposed a novel method for the prediction of individual-level risk scores from population-level interactome. We first constructed a DDN (disease-layer) which elaborates on the genetic associations among multiple phenotypes in UKBB PheWAS data. In order to use the disease-layer at the individual-level, we attached a SNP-layer to the disease-layer. The final developed network is a disease-SNP heterogeneous multi-layered network denoted as DS-Net. We employed graph-based SSL on the network to devise a network-based scoring algorithm. The SNP-layer is a single network that serves as initial labeling to receive individual genotyping data, and the disease-layer is an output network. The disease-layer serves as the predicted possible comorbidity risk scores in which the individual's genotype is propagated. To obtain layer-wise predicted scores, a layer-wise positive-unlabeled learning setting was employed, where the all nodes on the disease-layer are unlabeled and all the SNPs on the SNP-layer are labeled. Graph-based SSL can operate in this problem setting to propagate label information according to the topology of the network. The resulting netCRS is an estimated comorbidity score that integrates pre-defined genetic association between phenotypes using the underlying structure of the DS-Net. This score includes not only genetic information about a specific target disease, but also multiple associations of diseases. We validated the proposed netCRS by considering MI as index disease of interest. The netCRS model outperformed the conventional PRS-PT model in predicting MI patients and healthy controls. From experimental results of the association analysis, it is noteworthy that netCRS and PRS-PT work complementary to one another in identifying the very high-risk group of patients with myocardial infarction.

The current proposed method still has room for improvement. First, when constructing a disease-specific heterogeneous multi-layered network, it is expected that better comorbidity scores will be obtained if more precise criteria are applied to node selection. Second, our network was constructed using only common variants from PheWAS summary data. If we expand the network to include rare variants and other clinical information, we expect that using these risk prediction models will allow for the development of prevention strategies and reduction of MI morbidity and mortality. Also, the current disease-layer was constructed according to shared common SNPs between diseases. We can also try to build the DDN using different forms of genetic correlations such as LD regression scores. For future work, we will test netCRS in various diseases and compare netCRS with more recent PRS approaches in order to prove its generalized prediction performance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIGMS R01 GM138597, NLM R01 NL012535, and S10OD023495. We thank the staff of the Regeneron Genetics Center for DNA sequencing from PMBB participants. Use of the UK Biobank Resource in the current study was approved under Application Number 67855. Supplementary data are available at https://github.com/dokyoonkimlab/netCRS/blob/main/netCRS_supplemental.pdf

References

1. Choi SW, Mak TS-H & O'Reilly PF Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols 15, 2759–2772 (2020). [PubMed: 32709988]
2. Martin AR et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nature genetics 51, 584–591 (2019). [PubMed: 30926966]
3. Khera AV et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature genetics 50, 1219–1224 (2018). [PubMed: 30104762]
4. Elliott J et al. Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease. Jama 323, 636–645 (2020). [PubMed: 32068818]
5. Duncan L et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nature communications 10, 1–9 (2019).
6. Zheutlin AB et al. Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. American Journal of Psychiatry 176, 846–855 (2019).
7. Valderas JM, Starfield B, Sibbald B, Salisbury C & Roland M Defining comorbidity: implications for understanding health and health services. The Annals of Family Medicine 7, 357–363 (2009). [PubMed: 19597174]
8. Fritsche LG et al. Association of polygenic risk scores for multiple cancers in a phenome-wide study: results from the Michigan Genomics Initiative. The American Journal of Human Genetics 102, 1048–1061 (2018). [PubMed: 29779563]
9. Leppert B et al. A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK Biobank. PLoS genetics 16, e1008185 (2020). [PubMed: 32392212]
10. Meisner A et al. Combined utility of 25 disease and risk factor polygenic risk scores for stratifying risk of all-cause mortality. The American Journal of Human Genetics 107, 418–431 (2020). [PubMed: 32758451]
11. Goh K-I et al. The human disease network. Proceedings of the National Academy of Sciences 104, 8685–8690 (2007).
12. Nam Y et al. The translational network for metabolic disease–from protein interaction to disease co-occurrence. BMC bioinformatics 20, 1–12 (2019). [PubMed: 30606105]
13. Lee D-S et al. The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences 105, 9880–9885 (2008).
14. Verma A et al. Human-disease phenotype map derived from PheWAS across 38,682 individuals. The American Journal of Human Genetics 104, 55–64 (2019). [PubMed: 30598166]
15. Chong Y, Ding Y, Yan Q & Pan S Graph-based semi-supervised learning: A review. Neurocomputing 408, 216–230 (2020).
16. Zhou W et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics 50, 1335–1341 (2018). [PubMed: 30104761]
17. Denny JC et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 26, 1205–1210 (2010). [PubMed: 20335276]
18. Denny JC et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology 31, 1102–1111 (2013).

19. Wu P et al. Developing and Evaluating Mappings of ICD-10 and ICD-10-CM codes to Phecodes. BioRxiv, 462077 (2019).
20. Kember R et al. Polygenic Risk Scores for Cardio-renal-metabolic Diseases in the Penn Medicine Biobank. bioRxiv, 759381 (2019).
21. Choi SW & O'Reilly PF PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 8, giz082 (2019). [PubMed: 31307061]
22. Nikpay M et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature genetics 47, 1121 (2015). [PubMed: 26343387]
23. Lee B, Zhang S, Poleksic A & Xie L Heterogeneous multi-layered network model for omics data integration and analysis. Frontiers in genetics 10, 1381 (2020). [PubMed: 32063919]

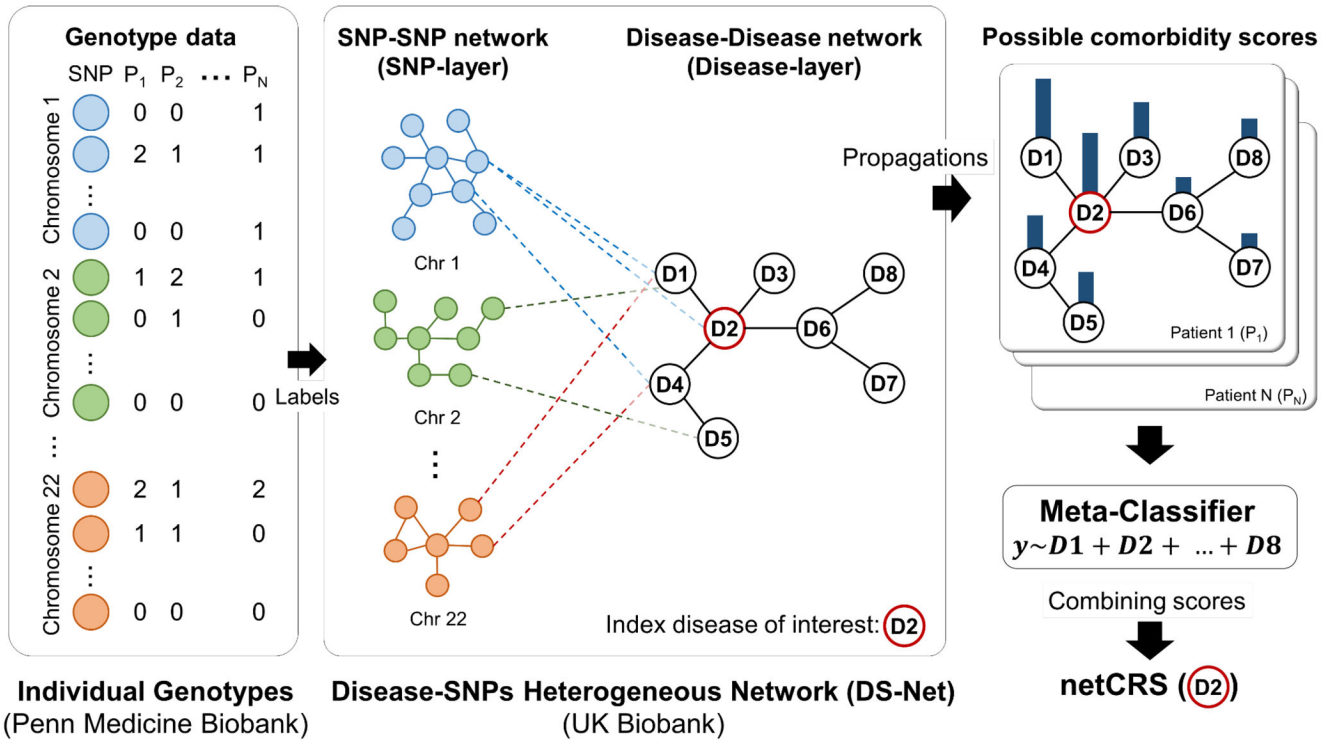

Figure 1. Overall framework of network-based comorbidity risk scoring algorithms (netCRS): **Left)** individual genotype data collected from Penn Medicine BioBank. **Middle)** schematic description of disease-SNP heterogeneous multi-layered network (DS-Net). SNP-layer constructed by linkage-disequilibrium and disease-layer constructed using UK biobank PheWAS summary data. **Right)** Upper right represents possible comorbidity scores of each disease for individual. The possible comorbidity scores are combined by logistic regression, and the combined scores, netCRS, are generated by each patient

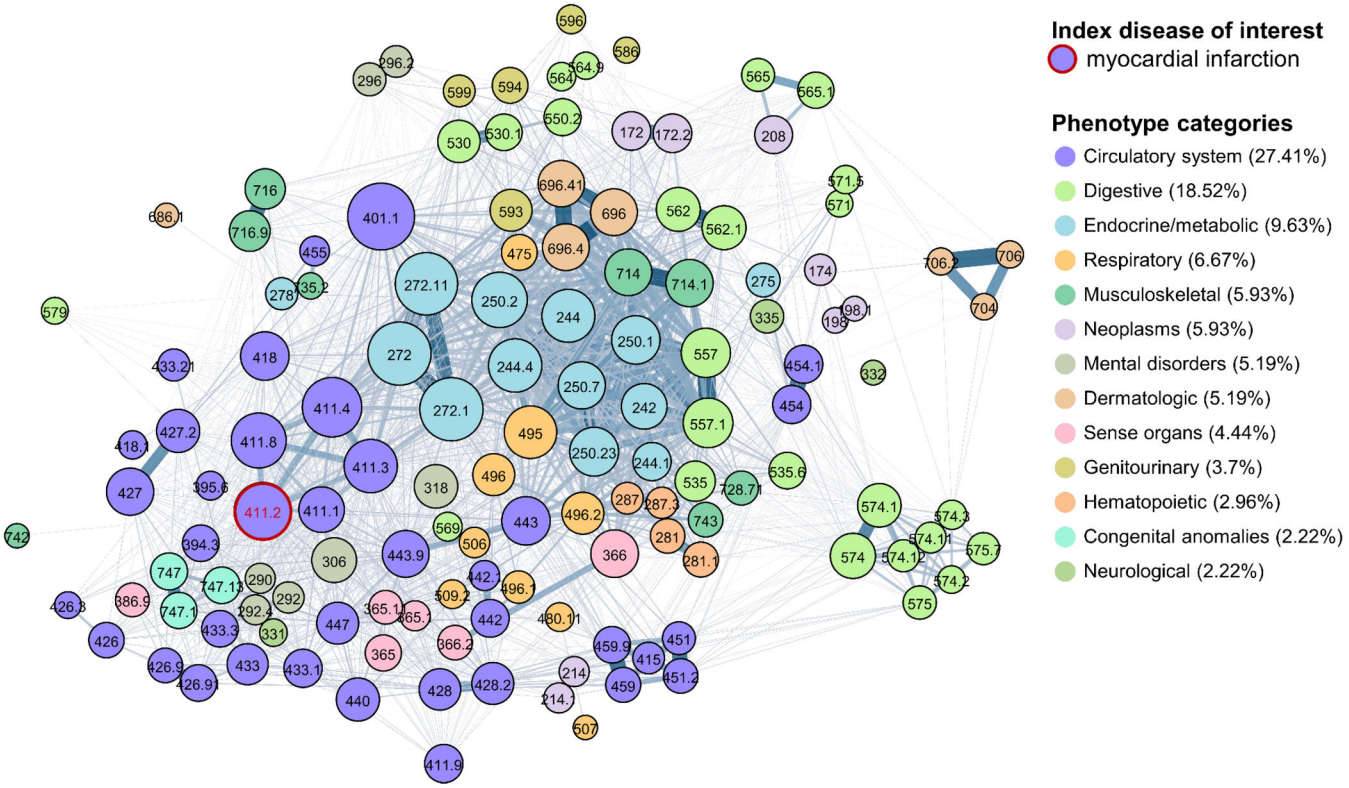

Figure 2. Visualization of MI-specific disease-layer:

The node size is the sum of the weighted degree of the node, indicating the relative size, and the node labels represents their PheCode. The thickness of the line represent the edge weights (similarity). Parentheses in disease categories represent the percentages of diseases that belong to a category.

Table 1.

Demographics table of the development and validation cohort.

Development Cohort (Network construction)		***UK BioBank PheWAS summary data (UKBB)***			
	Phenotypes	**135 (out of 1,403)**			
	SNPs	**39,365 (after genetic pre-processing)**			
		Penn Medicine BioBank (PMBB)			
		Total	**MI cases**	**Controls**	***p*-value**
Validation Cohort (Genotype data)	**No. of samples**	(N = 4972)	(N = 763)	(N = 4209)	
	Sex				<0.001
	Female (%)	1,854 (37.3%)	171 (22.4%)	1683 (40.0%)	
	Male (%)	3,118 (62.7%)	592 (77.6%)	2526 (60.0%)	
	Age at enrollment	62.0 ± 14.8	68.4 ± 11.2	60.9 ± 15.1	<0.001

Table 2.

Performance comparison of netCRS and PRS-PT in terms of AUC

Models	Hyper-parameter (μ) for netCRS				
	0.01	**0.1**	**1**	**10**	**100**
[1] PRS-PT	0.5827 (Baseline)				
[2] netCRS*	0.6028	**0.6444**	0.6395	0.6197	0.6039
[3] netCRS + PRS-PT	0.6274	**0.6609**	0.6570	0.6389	0.6255
[4] PRS-PT + Sex + Age	0.6979 (Baseline)				
[5] netCRS + Sex + Age*	0.7083	**0.7287**	0.7261	0.7144	0.7051
[6] netCRS + PRS-PT + Sex + Age*	0.7230	**0.7417**	0.7396	0.7287	0.7199

Table 3.

Diagnostic odds ratio and 95% confidential intervals for the MI according to netCRS risk group: We compared three different models: (a) model [2]: netCRS alone, (b) model [5]: netCRS + sex + age, and (c) model [6]: netCRS + PRS-PT + sex + age.

Total (N = 4,972)	No. of MI/ No. of Total	Model [2]		Model [5]		Model [6]	
		OR (95% CI)	*p*-value*	OR (95% CI)	*p*-value*	OR (95% CI)	*p*-value*
Low risk (0^{th}-25^{th})	94/1243	Reference					
Intermediate risk (26^{th}-50^{th})	150/1243	1.68 (1.28–2.21)	<0.001	1.71 (1.30–2.25)	<0.001	1.65 (1.25–2.19)	<0.001
High risk (51^{st}-75^{th})	218/1243	2.60 (2.02–3.37)	<0.001	2.72 (2.10–3.55)	<0.001	2.70 (2.08–3.53)	<0.001
Very high risk (76^{th}-100^{th})	301/1243	**3.91 (3.06–5.02)**	**<0.001**	**4.01 (3.13–5.50)**	**<0.001**	**3.83 (2.98–4.96)**	**<0.001**

Abbreviations: OR, odds ratio; CI, confidence interval; PRS, polygenic risk score.

** p*-value for netCRS categories.

Table 4.

Genetic subgroups based on the combinations of PRS and netCRS

	Odds ratio* **(No. of MI / No. of Total)**	**PRS-PT(MI)**			
		Low risk (0^{th}-25^{th})	**Intermediate risk (26^{th}-50^{th})**	**High risk (51^{st}-75^{th})**	**Very high risk (76^{th}-100^{th})**
netCRS(MI)	Low risk (0^{th}-25^{th})	Reference (19/334)	1.18 (20/299)	1.35 (21/273)	2.46 (34/243)
	Intermediate risk (26^{th}-50^{th})	1.46 (23/276)	2.36 (36/268)	2.77 (45/286)	3.07 (46/263)
	High risk (51^{st}-75^{th})	2.07 (33/280)	4.59 (71/272)	3.94 (52/241)	4.55 (60/232)
	Very high risk (76^{th}-100^{th})	4.04 (52/226)	4.66 (58/219)	5.60 (78/245)	**7.88 (113/252)**

*For calculating odds ratio, we performed multivariate logistic regression analysis for MI classification task (myocardial infarction (MI) cases versus Normal control). Logistic model: (MI cases vs. Normal control) ~ 16 combinations (PRS and netCRS groups) + sex + age. With the lowest risk group (Low PRS group & Low netCRS group) as a reference, the odds ratio of each combination was reported in this table.