MorbidGCN: prediction of multimorbidity with a graph convolutional network based on integration of population phenotypes and disease network

Guifying Dong, Zi-Chao Zhang, Jianfeng Feng and Xing-Ming Zhao

Corresponding author. Xing-Ming Zhao, Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China; MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, and MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China; Zhangjiang Fudan International Innovation Center, Shanghai 200433, China. International Human Phenome Institutes (Shanghai), Shanghai 200433, China. Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China. Tel.: 021-55665546; E-mail: xmzhao@fudan.edu.cn

Abstract

Exploring multimorbidity relationships among diseases is of great importance for understanding their shared mechanisms, precise diagnosis and treatment. However, the landscape of multimorbidities is still far from complete due to the complex nature of multimorbidity. Although various types of biological data, such as biomolecules and clinical symptoms, have been used to identify multimorbidities, the population phenotype information (e.g. physical activity and diet) remains less explored for multimorbidity. Here, we present a graph convolutional network (GCN) model, named MorbidGCN, for multimorbidity prediction by integrating population phenotypes and disease network. Specifically, MorbidGCN treats the multimorbidity prediction as a missing link prediction problem in the disease network, where a novel feature selection method is embedded to select important phenotypes. Benchmarking results on two large-scale multimorbidity data sets, i.e. the UK Biobank (UKB) and Human Disease Network (HuDiNe) data sets, demonstrate that MorbidGCN outperforms other competitive methods. With MorbidGCN, 9742 and 14 010 novel multimorbidities are identified in the UKB and HuDiNe data sets, respectively. Moreover, we notice that the selected phenotypes that are generally differentially distributed between multimorbidity patients and single-disease patients can help interpret multimorbidities and show potential for prognosis of multimorbidities.

Keywords: multimorbidity, graph convolutional network, population phenotype, disease network

Introduction

Multimorbidity refers to the co-occurrence of multiple diseases in a patient not by chance [1], which has greatly increased global health burdens, as people suffering from multimorbidity tend to have higher mortality rate and medical expenditure [2, 3]. Nevertheless, the known multimorbidity relationships, mainly inferred from electronic health records (EHRs), are far from complete, as the available EHRs usually have small sample sizes, short follow-up duration, biased population structures and missed diagnoses [4–6]. For example, the EHRs in the UK Biobank (UKB) only involve about 500 000 individuals aged between 40 and 69, while the UKB population is biased to be healthy [4]. As another example, the EHRs in the Human Disease Network (HuDiNe) only have a 3-year follow-up duration with all patients over 65 years old, although it has a large sample size (13 039 018 individuals) [5]. Therefore, a computational method to predict the missing multimorbidities is necessary to supplement the known multimorbidities from EHRs.

It is challenging to predict multimorbidities, as human diseases are usually associated with multiple factors due to their heterogeneity and polygenicity, especially for complex diseases [7]. Nevertheless, much effort has been made to investigate the multimorbidity relationships among diseases by quantifying disease similarities based on different types of biological data. Biomolecular networks, such as protein interaction networks and gene networks, are the most common biological data that were used to quantify disease similarities for multimorbidity study [8–11]. For example, Menche et al. [10] calculated disease similarities in a protein interaction network and found that similar diseases showed significant multimorbidity patterns. Li et al. [11] developed a deep representation learning model to learn disease representations from a gene network and found that disease similarities calculated based on the learned representations could better predict disease–disease relationships. In addition, clinical symptoms were also used to quantify disease similarities, and diseases with
high symptom-based similarity exhibited significant multimorbidity, such as Alzheimer’s disease and epilepsy [12]. Moreover, several studies also integrated omics data to define disease similarities [13, 14]. For example, Oerton et al. [14] calculated disease similarities by integrating six different types of biological data, including ontology, clinical symptom, literature co-occurrence, genetic association, gene expression and drug indication. They found that similar diseases had higher co-occurrence frequencies in the same patient. These results demonstrated that disease similarities quantified based on biological data can effectively characterize multimorbidity relationships. However, due to the limited knowledge about the biological mechanisms of multimorbidities, these methods can only interpret a small number of multimorbidities.

Compared with biomolecules and literature-mined terms, phenotypes among the general population (e.g. physical activity and diet) have been neglected previously, as they are rarely collected together with disease diagnoses, although they can characterize a person more comprehensively. In this study, we called such phenotypes as population phenotypes. It is well recognized that population phenotypes can contribute to the coexistence of multiple diseases in a patient. For example, smoking has been reported to increase the risk of hypertension, chronic obstructive pulmonary disease, gastroesophageal reflux disease and chronic pharyngitis for patients with obstructive sleep apnea [15]. Besides, population phenotypes, as one type of symptoms, are usually associated with multimorbidities [16]. In addition, disease networks have been commonly used to reveal hidden connections among biomedical entities such as diseases, drugs and genes [17–19]. Lee et al. [20] found that connected diseases in the disease network constructed by linking diseases with correlated metabolic reactions displayed high multimorbidity patterns. Therefore, disease networks can possibly be used to reveal hidden multimorbidity relationships among diseases. In order to learn useful node representations from networks, many algorithms have been proposed, such as DeepWalk [21], Node2Vec [22] and graph convolutional network (GCN) [23–25]. Compared with other methods, GCN is more advantageous in integrating node features and network topology and has been shown to outperform other methods in a variety of tasks, such as node classification [23, 26] and link prediction [24].

Considering the above, we introduced MorbidGCN, a GCN method that integrates population phenotypes and disease network for multimorbidity prediction. MorbidGCN treated the multimorbidity prediction problem as a missing link prediction task in the disease network, where a novel feature selection method was embedded to select important phenotypes. Benchmarking results on two large-scale multimorbidity data sets demonstrated the improved prediction performance of MorbidGCN compared with other methods. With MorbidGCN, tens of thousands of novel multimorbidities were discovered. Moreover, we found that the selected phenotypes tended to be differentially distributed between multimorbidity patients and single-disease patients and could help prognose the multimorbidity occurrence. MorbidGCN can be easily applied on other EHRs to predict the missing multimorbidities for certain populations.

Materials and methods
Multimorbidity data sets
In this study, the multimorbidity was defined by two measures, i.e. relative risk (RR) and P-value (significance of co-occurrence), which have been used in previous studies [1, 27]. Accordingly, two multimorbidity data sets, i.e. the UKB and HuDiNe data sets, were used, as the RR and P-value for each disease pair in these data sets can be obtained.

The multimorbidity relationships in the UKB data set were derived from the fields related to the first occurrence date of diseases in the UKB (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=1712). The diseases in these fields were defined by the 3-character International Classification of Disease version 10 (ICD10) codes. For each disease pair, we calculated its RR and P-value [1]. Disease pairs with RR > 1 and P-value < 0.05 (Bonferroni corrected) were taken as multimorbidities.

The multimorbidity relationships in the HuDiNe data set were derived from the summary statistics of disease pairs provided by Hidalgo et al. [5], including \(I_i\), (number of patients with the disease \(i\)), \(I_j\) (number of patients with the disease \(j\)), \(C_{ij}\) (number of patients with both the diseases \(i\) and \(j\)) and RR \(RR_{ij}\) (http://sbi.upf.edu/data/hudine/). To obtain the HuDiNe multimorbidities, we first calculated the P-value for each disease pair based on the reported \(I_i\), \(I_j\) and \(C_{ij}\) [1]. Then, disease pairs with RR > 1 and false discovery rate (FDR)-adjusted P-value < 0.05 were taken as multimorbidities. Here, a relatively loose multiple test correction method, that is, the FDR correction, was used. This is because the follow-up duration (3 years) of the HuDiNe diagnosis data is much shorter than that (decades) of the UKB diagnosis data, which can result in losing power to detect multimorbidities. Finally, as the diseases in HuDiNe are the 3- and 5-digit ICD9 codes, we mapped them to the 3-character ICD10 codes.

For both data sets, only the diseases with phenotype scores (described in the following part) were considered. As a result, 52 670 multimorbidity disease pairs among 573 diseases were found in the UKB data set (Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/), and 41 359 multimorbidity disease pairs among 494 diseases were found in the HuDiNe data set (Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/).
Evaluation metrics

We used the area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPRC) to evaluate the multimorbidity prediction performance.

Overview of MorbidGCN

MorbidGCN integrates population phenotypes and disease network based on GCN for multimorbidity prediction. The workflow of MorbidGCN includes four steps. (i) Phenotype processing and encoding: each population phenotype was processed to form a single valid value per participant and was encoded into either continuous, binary or ordered categorical variables. (ii) Disease-phenotype association quantification: three kinds of correlation coefficients were used to quantify the associations between diseases and different types of phenotypes. (iii) Phenotype selection: as the dimension of the population phenotypes is very high, we proposed a novel feature selection method to select useful phenotypes. (iv) GCN for multimorbidity prediction: we constructed a disease network based on the known multimorbidity relationships and casted the multimorbidity prediction problem as a missing link prediction task in the disease network. A two-layer GCN was used to integrate the selected phenotypes with the disease network for multimorbidity prediction. The above four steps were detailed as follows.

Phenotype processing and encoding

The 2372 phenotypes in the UKB were preselected and were divided into 18 groups, i.e. physical measures, blood, urine, brain magnetic resonance imaging (MRI), heart MRI, abdominal MRI, dual-energy X-ray absorptiometry (DXA) assessment, diet, physical activity, sleep, smoking, sun exposure, electronic device use, local environment, household, employment, deprivation and early life factors (Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). To process and encode phenotypes, the following three steps were performed.

Step 1. Preprocessing phenotypes before encoding

The invalid values of phenotypes (e.g. the value ‘−3′ which represents ‘Prefer not to answer’ for field ‘20116: Smoking status’) were removed. For phenotypes measured with more than one times, only the first value was used per participant.

Step 2. Phenotype encoding

The values of the phenotypes from UKB are of numerical or categorical types, and the details about the value types of phenotypes can be found on the webpage of UKB (https://biobank.ctsu.ox.ac.uk/showcase/list.cgi). We encoded these phenotypes as either continuous, binary or ordered categorical variables according to their value types (Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). Specifically, the phenotypes were encoded as below.

(i) For phenotypes with numerical values, we directly used their values and encoded them as continuous variables.

(ii) For phenotypes with only two categorical values, we encoded them as binary variables. For example, the phenotype ‘100580-Alcohol consumed’ only has two values – ‘1: Yes’ and ‘0: No’. For this phenotype, its values were encoded as either 1 or 0, where 1 denotes participants consume alcohol (for Yes) and 0 denotes participants do not consume alcohol (for No).

(iii) For phenotypes with more than two categorical values:

(a) If their values have intrinsic orders, we encoded them as ordered categorical variables. For example, the phenotype ‘22606-Workplace very noisy’ has three values that can be sorted in the frequency order of ‘1: Often’, ‘2: Sometime’ and ‘3: Rarely/never’. The values of this phenotype were encoded as either 0, 1 or 2, with 0 for workplace often noisy, 1 for sometimes noisy and 2 for rarely/never noisy.

(b) If their values have no order attributes, we encoded each value of these phenotypes as a binary variable. For example, the phenotype ‘6143-Transport type for commuting to job workplace’ with four possible values—‘1: Car/motor vehicle’, ‘2: Walk’, ‘3: Public transport’ and ‘4: Cycle’, it will be encoded as (0, 1, 0, 0) if the value is ‘Walk’.

Step 3. Postprocessing phenotypes after encoding

Continuous phenotypes were transformed into Z-scores. Binary and ordered categorical phenotypes with extremely unbalanced proportions of values were removed, i.e. the binary phenotype with its most common value accounting for more than 99% participants and the ordered categorical phenotype with the frequency of its second common value lower than 5% frequency of its most common value among participants.

Quantifying disease–phenotype associations based on the UKB population

Three types of correlation coefficients, i.e. the point bivariate correlation coefficient ($r_{b}$), the phi correlation coefficient ($r_{p}$) and the Kendall’s tau ($\tau$) [28], were used to quantify the disease–phenotype associations for continuous, binary and ordered categorical phenotypes, respectively (for details, see Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). With the quantification, we obtained a disease-phenotype association matrix ($X_{c} \in \mathbb{R}^{m \times n}$), with $m$ rows indexed by diseases, $n$ columns indexed by phenotypes and the values being the correlation coefficients between diseases and phenotypes. The disease–phenotype pairs with overlaps of no more than 10 patients were not
considered, which results in missing values in \( X \). To process missing values, we first removed the diseases with phenotype missing rate \( \geq 10\% \) and then removed the phenotypes with disease missing rate \( \geq 10\% \). Afterward, the missing values were imputed group by group. Specifically, for each disease, its missing phenotype values were imputed using the mean values of its five nearest neighbors found in the corresponding phenotype group. By doing so, the phenotype information in the same group can be effectively utilized, ensuring more precise imputation.

Finally, the associations between 2382 phenotypes and 573 diseases were obtained. Here, the phenotypes mean encoded phenotypes, where some phenotypes with more than two categorical values were encoded into several binary phenotypes during phenotype encoding, leading to the difference between the number of encoded phenotypes (2382) and the initially selected phenotypes (2372).

**Phenotype selection**

A novel feature selection method was proposed to select important phenotypes that can better distinguish multimorbidities from non-multimorbidities. First, we flattened \( X_c \in \mathbb{R}^{m \times n} \) into a new disease–phenotype matrix \( X \in \mathbb{R}^{m \times 2n} \). The values in the first \( n \) columns of \( X \) are the positive correlation coefficients with all negative coefficients replaced by 0, and the values in the last \( n \) columns are the absolute values of the negative correlation coefficients with all positive coefficients replaced by 0. Second, \( X \) was normalized by the min–max normalization. Third, we used a diagonal matrix \( W \in \mathbb{R}^{2n \times 2n} \), i.e. \( W = \text{diag}(w_1, w_2, \ldots, w_{2n}) \) to weight phenotypes in \( X \). The similarities of disease pairs were measured by the inner product of their weighted phenotype vectors, and the disease similarity matrix \( S \in \mathbb{R}^{m \times m} \) was expressed as

\[
S = XW(XW)^T. \tag{1}
\]

Finally, we used the sigmoid function (\( \sigma \)) to convert the similarity scores into the predicted multimorbidity probabilities (\( \tilde{Y} \in \mathbb{R}^{m \times m} \)), i.e.

\[
\tilde{Y} = \sigma(S) = \frac{1}{1 + e^{-XW(XW)^T}}. \tag{2}
\]

In order to select useful phenotypes, we minimized the discrepancy between the true (\( Y \)) and predicted (\( \tilde{Y} \)) multimorbidity relationships based on the cross-entropy loss function as shown below

\[
\min_{\tilde{Y}} \mathcal{L}(W) = -\frac{1}{m^2 - m} \sum_{i=1}^{m} \sum_{j=1}^{m} Y_{ij} \log \left( \tilde{Y}_{ij} \right) + (1 - Y_{ij}) \log \left( 1 - \tilde{Y}_{ij} \right), \tag{3}
\]

where \( Y_{ij} \) and \( \tilde{Y}_{ij} \) denote the true label and the predicted multimorbidity probability between diseases \( i \) and \( j \), respectively. \( \tilde{Y}_{ij} \) can be written as

\[
\tilde{Y}_{ij} = \sigma \left( S_{ij} \right) = \sigma \left( \sum_{k=1}^{2n} w_{ik} x_i x_k \right). \tag{4}
\]

However, as \( \mathcal{L}(W) = \mathcal{L}(-W) \), minimizing \( \mathcal{L} \) [Equation (3)] with respect to \( W \) is a nonconvex optimization problem. To convert this problem into a convex optimization problem, we defined a vector \( Z \in \mathbb{R}^{2n} \), i.e.\( Z = [w_1, w_2, w_3, \ldots, w_{2n}]^T \), with which the predicted multimorbidity probability between diseases \( i \) and \( j \) [Equation (4)] can be converted to

\[
\tilde{Y}_{ij} = \sigma \left( \sum_{k=1}^{2n} z_k x_i x_k \right), \text{ s.t.} z_k \geq 0, \text{ for } k = 1, \ldots, 2n. \tag{5}
\]

and the objective function [Equation (3)] can be converted to

\[
\min_Z \mathcal{L}(Z) = -\frac{1}{m^2 - m} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} Y_{ij} \log \left( \tilde{Y}_{ij} \right) + (1 - Y_{ij}) \log \left( 1 - \tilde{Y}_{ij} \right), \text{ s.t.} Z \geq 0. \tag{6}
\]

Actually, Equation (6) is the objective function of the logistic regression with a positive semidefinite constraint on \( Z \), so it is a convex function. Furthermore, in order to achieve sparse feature selection, L1 regularization was used. The final objective function is defined as follows.

\[
\min_Z \mathcal{L}(Z) = -\frac{1}{m^2 - m} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} Y_{ij} \log \left( \tilde{Y}_{ij} \right) + (1 - Y_{ij}) \log \left( 1 - \tilde{Y}_{ij} \right) + \lambda \| Z \|_1, \text{ s.t.} Z \geq 0. \tag{7}
\]

We used the gradient descent algorithm to minimize Equation (7). The updating rule for \( Z \) is

\[
Z = Z - \frac{\partial \mathcal{L}(Z)}{\partial Z}, \quad z_k = 0 \text{ for } z_k < 0, \tag{8}
\]

where \( \alpha \) is the learning rate, and

\[
\frac{\partial \mathcal{L}(Z)}{\partial Z} = -\frac{1}{m^2 - m} \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} \left( Y_{ij} - \frac{1}{1 + e^{-x_i \cdot x_j}} \right) x_i (x_i \cdot x_j)^T + \lambda \text{sign}(Z). \tag{9}
\]

Here, \( x_i \) and \( x_j \) are the \( i \)th and \( j \)th rows of \( X \), respectively. If \( z_k = 0 \), \( \text{sign}(z_k) = 0 \).

With the above approach, phenotypes with their corresponding \( Z \) elements greater than a certain threshold (\( \epsilon \)) were selected. The training and evaluation of the feature
selection method are described in Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/.

GCN for multimorbidity prediction
We constructed a disease network with nodes being diseases and edges being the known multimorbidity relationships. The adjacency matrix of the network is denoted as $A \in \mathbb{R}^{m \times m}$. Each value in $A$ is either 1 or 0, with 1 indicating the corresponding two diseases are the known multimorbidity and 0 indicating they are not.

MorbidGCN used the two graph convolution layers to integrate the population phenotypes ($X$) and disease network ($A$) for multimorbidity prediction. The first and second layers were individually defined as follows:

$$\hat{H}^{(1)} = f \left( X, \hat{A} \right) = \text{ReLU} \left( \hat{A} X W^{(0)} \right),$$

$$\hat{H}^{(2)} = f \left( \hat{H}^{(1)}, \hat{A} \right) = \hat{A} \hat{H}^{(1)} W^{(1)},$$

where $\hat{A} = \hat{D}^{-1/2} (A + I) \hat{D}^{-1/2}$, $\hat{D}$ is the diagonal degree matrix of $(A + I)$, and $I$ is the identity matrix. $W^{(0)}$ and $W^{(1)}$ are the weight matrices of the first and second layers, respectively. The output ($\hat{H}^{(2)}$) of the second layer was taken as the final disease embeddings, which was used to reconstruct the adjacency matrix ($\hat{A}$) of the disease network, that is

$$\hat{A} = \delta \left( \hat{H}^{(2)} \left( \hat{H}^{(2)} \right)^T \right).$$

The loss function is

$$L = -\frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} A_{ij} \log \left( \hat{A}_{ij} \right) + (1 - A_{ij}) \log \left( 1 - \hat{A}_{ij} \right),$$

where $A_{ij}$ and $\hat{A}_{ij}$ are the true label and the predicted multimorbidity probability between diseases $i$ and $j$, respectively.

MorbidGCN was trained on an incomplete version of the disease network where parts of the edges were removed. The removed edges combined with the same number of randomly sampled pairs of unconnected nodes (non-edges) formed the positive and negative sets for evaluation usages. We evaluated whether the disease embeddings learned from parts of the known multimorbidity relationships could be used to predict the missing multimorbidities (for details, see Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Results
Framework of MorbidGCN
The architecture of MorbidGCN is depicted in Figure 1, where four steps were performed. First, 2372 phenotypes in the UKB were preselected (Figure 1A) and were processed and encoded into either continuous, binary or ordered categorical types (Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). Second, the relationships between phenotypes and diseases were quantified based on the UKB population, where three kinds of correlation coefficients were used to respectively quantify the disease-phenotype associations for continuous, binary and ordered categorical phenotypes (Figure 1B). Missing scores were imputed by the group k nearest neighbor (KNN) method (Materials and Methods). Third, as the dimension of population phenotypes is very high, a novel feature selection method was proposed (Figure 1C). The method weights each feature by a coefficient which indicates feature’s importance. The weight coefficients were optimized to ensure that the disease similarities calculated based on the weighted features can distinguish the known multimorbid disease pairs from other disease pairs to the largest extent. Phenotypes with weight coefficients greater than a certain threshold were selected. Finally, we constructed a disease network based on known multimorbidity relationships and regarded the multimorbidity prediction as a missing link prediction task in the network. A two-layer GCN was used to integrate the selected phenotypes with the disease network for multimorbidity prediction (Figure 1D). We benchmarked our model on two multimorbidity data sets, i.e. the UKB and HuDiNe data sets [5] (Supplementary Tables S1 and S2, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Selected phenotypes can better predict multimorbidities
MorbidGCN includes a novel feature selection step, which was used to select useful phenotypes for multimorbidity prediction (see Materials and Methods). First, we tested whether the selected phenotypes were affected by the data set partition. Specifically, we performed feature selection 100 times, and for each time, the data set was randomly split into training (90%) and test (10%) sets. The training set was used for phenotype selection (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). We found that the number of the selected phenotypes was almost the same under different data set partitions for both the UKB and HuDiNe data sets (Supplementary Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). Moreover, the selected phenotypes were highly repeatable, with about 91% and 80% of the selected phenotypes being repeatable among more than half of the data set partitions for the UKB and HuDiNe data sets, respectively (Figure 2A, B). Then, we ranked the selected phenotypes by their weight coefficients and calculated their rank similarities under different data set partitions by the rank-biased overlap (RBO) [29] (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). As shown in Figure 2C, D, the RBO scores were very
Figure 1. Framework of MorbidGCN. (A) Processing and encoding phenotypes in the UKB. (B) Quantification of disease–phenotype associations and imputation of missing values. (C) Feature selection method for multimorbidity prediction. (D) Two-layer GCN integrating the selected phenotypes with a disease network for multimorbidity prediction.

High, which mainly distributed in the range from 0.96 to 0.97 for the UKB data set and from 0.94 to 0.98 for the HuDiNe data set, suggesting the proposed feature selection method can rank phenotypes stably.

Second, we evaluated whether the disease similarities calculated based on the selected phenotypes could better distinguish the known multimorbid disease pairs from other disease pairs than those calculated
Figure 2. Selected phenotypes among 100 randomly generated data set partitions of the UKB and HuDiNe data sets. Repeatability of the selected phenotypes among the 100 partitions of the UKB (A) and HuDiNe (B) data sets. The label in each slice of a pie chart represents the range of repetition times. The percentage in each slice of a pie chart represents the proportions of the selected phenotypes that are repeatable in certain times. Rank similarities of the selected phenotypes under the 100 partitions of the UKB (C) and HuDiNe (D) data sets. Multimorbidity prediction performance based on the selected phenotypes, all phenotypes and randomly selected phenotypes (the same number of phenotypes was selected from all phenotypes and from the remaining phenotypes except those selected by our feature selection method) under the 100 partitions of the UKB (E) and HuDiNe (F) data sets. The $P$ denotes $P$-value generated by the two-sided Mann–Whitney U test. The maximum weights of the selected phenotypes among the 100 partitions of the UKB (G) and HuDiNe (H) data sets.

Based on (i) all phenotypes, (ii) phenotypes randomly selected from all phenotypes and (iii) phenotypes randomly selected from the remaining phenotypes except those selected by our method (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). This evaluation was performed on the independent test sets formed by the above-mentioned 100 data set partitions. As shown in Figure 2E, F, the prediction performance based on the selected phenotypes was significantly improved.
compared with that based on all phenotypes (the average AUC and AUPRC were improved by 8.3 and 13.5%, for the UKB data set and 11.1 and 7.8% for the HuDiNe data set, respectively). In contrast, when the phenotypes were randomly selected from either all phenotypes or the remaining phenotypes, the prediction performance was significantly decreased compared with that of all phenotypes (Figure 2E, F). Moreover, we also compared the proposed feature selection method with two commonly used feature selection methods, i.e., logistic regression with the L1 penalty (denoted as LR-L1) and random forest (denoted as RF). We found that the features selected by our method could better predict multimorbidities than the features selected by other methods (Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/).

We defined the final selected phenotypes as those that had been selected at least once among the 100 data set partitions. A total of 237 and 78 phenotypes were selected on the UKB and HuDiNe data sets, respectively (Supplementary Tables S4 and S5, see Supplementary Data available online at http://bib.oxfordjournals.org/). The selected phenotypes on the two data sets were highly overlapped (55 phenotypes overlapped). Figure 2G, H shows the maximum weights of the selected phenotypes under different data set partitions. We found a wide range of phenotypes associated with multimorbidities, among which multiple blood phenotypes had high weights on both data sets, for example, fields of 30720, 30770, 30260, 30180, 30200 and 30070. This is consistent with previous findings that these blood biomarkers were associated with multiple diseases, such as coronary heart disease [30], sepsis [31], acute respiratory distress syndrome [32], chronic kidney disease [33] and the multimorbidity of heart failure and diabetes [34]. In addition, our results also highlighted the effects of several smoking phenotypes (e.g. fields of 2887, 20161, 3476, 3456 and 3466) on multimorbidities, which is supported by previous conclusions that cigarette smoking harmed almost every human organ and remained a leading cause of preventable diseases [35].

Population phenotypes can better predict multimorbidities than other biological data
MorbidGCN is the first work that uses population phenotypes for multimorbidity prediction. Therefore, we evaluated whether the disease similarities calculated based on the selected phenotypes could better predict multimorbidities than those calculated based on other types of biological data [10–12, 14]. Four methods—sAB [10], GeneNetRR [11], HSDN [12] and FusedSim [14] were compared with our phenotype-based method. The sAB, GeneNetRR, HSDN and FusedSim calculated disease similarities based on the protein interaction network, gene network, clinical symptoms and multi-omics data fusion (i.e. ontology, clinical symptom, literature occurrence, genetic association, gene expression and drug data), respectively (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). As shown in Figure 3A–D, our phenotype-based method achieved higher AUC and AUPRC than the other four methods on both the UKB and HuDiNe data sets. Among the four compared methods, FusedSim had the smallest performance gap with our method (Figure 3D), suggesting that integration of multi-omics data could improve the performance of multimorbidity prediction. However, due to the limited knowledge of disease–biomolecule associations, FusedSim only focused on 84 diseases [14], while our method can predict multimorbidity relationships among more than 500 diseases.

In particular, population phenotypes are more useful in predicting cross-physiological multimorbidities than biomolecules and literature-mined terms. As shown in Figure 3E, F, the four compared methods [10–12, 14] showed poor performance in predicting cross-physiological multimorbidities, where all AUCs and AUPRCs for cross-physiological multimorbidities were less than 0.6 on both data sets. On the other hand, our method had better performance in predicting both the same and cross-physiological multimorbidities, with all AUCs and AUPRCs greater than 0.83 on the UKB data set and most AUCs around 0.75 on the HuDiNe data set (Figure 3E, F).

MorbidGCN integrating disease network with population phenotypes improves prediction performance
We examined the necessity of different parts of MorbidGCN (i.e. the GCN, selected phenotypes and disease network) for multimorbidity prediction by extensive ablation studies, where five methods were compared, i.e. MorbidGCN−, MorbidGCN+, MorbidPhe, DeepWalk [21] and Node2Vec [22] (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). MorbidGCN− and MorbidGCN+ are two variant GCN models, which predict multimorbidities with only the disease network information and the integration of the disease network with all phenotypes, respectively. MorbidPhe predicts multimorbidities only based on phenotype information. DeepWalk and Node2Vec are two other network representation learning methods, which only use the disease network information for multimorbidity prediction. These methods were compared on 100 different randomly generated training, validation and test splits of data sets. As shown in Figure 4A, B, MorbidGCN outperformed all other methods on both data sets (the average AUC and AUPRC were 94.0 and 94.5% on the UKB data set and were 87.1 and 87.9% on the HuDiNe data set, respectively). On the one hand, MorbidGCN had better performance than MorbidGCN+ (Figure 4A, B), confirming that the proposed feature selection method could pick out important phenotypes while reducing the impact of useless phenotypes for multimorbidity prediction. On the other hand, MorbidGCN outperformed...
Novel multimorbidities identified by MorbidGCN

With MorbidGCN, 9742 and 14,010 novel multimorbidities were discovered on the UKB and HuDiNe data sets, respectively (Supplementary Methods; Supplementary Tables S6 and S7, see Supplementary Data available online at http://bib.oxfordjournals.org/). To validate the discovered multimorbidities, we performed enrichment analysis of the novel multimorbidities of the UKB data set in the known multimorbidities from the HuDiNe data set and vice versa (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). We found that the novel discovered multimorbidities from the two data sets were able to be mutually verified. The 2466 (25.31%) novel UKB multimorbidities were overlapped with the known HuDiNe multimorbidities, and this overlap was much higher than random (P-value < 1e−4, Figure 4C). Similarly, the novel HuDiNe multimorbidities were significantly overlapped with the known UKB multimorbidities (4422 overlapped multimorbidities; P-value < 1e−4, Figure 4C). In addition, 1626 novel multimorbidities were identified on both data sets, also being significantly higher than random (P-value < 1e−4, Figure 4C). These results confirmed the reliability of the novel discovered multimorbidities.

Interestingly, we found that MorbidGCN discovered a higher proportion of cross-physiological multimorbidities than expected (P-value < 1e−4; see Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/), with about 96% of the novel identified multimorbidities being the cross-physiological ones in both data sets (Supplementary Tables S6 and S7, see Supplementary Data available online at http://bib.oxfordjournals.org/). For example, the disease pair of 'E66-Obesity' and 'M81-Osteoporosis without pathological fracture' was a novel cross-physiological multimorbidity with the highest predicted probability in the UKB data set (Supplementary Table S6, see Supplementary Data available online at http://bib.oxfordjournals.org/), which can be supported by previous findings that the incidence of obesity was increased on the osteoporosis patients with pharmacological interventions [36], and the increasing of fat mass was negatively associated with the bone mass [37]. In addition, the novel multimorbidities had significantly higher topological similarity and phenotypic similarity (Supplementary Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/) compared with the non-multimorbidities, indicating the integration of phenotypes and disease network is indeed useful for multimorbidity prediction. Moreover, we identified some novel multimorbid disease pairs that cannot be detected statistically from EHRs even those disease pairs co-occur with high frequency (i.e. RR > 1). For example, 'J17-pneumonia in diseases classified elsewhere' and 'H54-blindness and low vision' had a high RR of 2.3 with P-value of 0.052 based on the EHRs from the UKB. Therefore, the two diseases were not a multimorbidity based on the information from EHRs. With MorbidGCN, we predicted this pair of diseases to be a multimorbidity that has actually been reported to be a multimorbidity in previous studies [38, 39].

The selected phenotypes tend to be differentially distributed between multimorbidity patients and single-disease patients

We evaluated whether the selected phenotypes were differentially distributed between multimorbidity/double-disease patients and single-disease patients (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). As shown in Figure 5A, there was about 1.1% of known multimorbidities, whose patients were significantly different from single-disease patients at each selected phenotype, while this proportion was only about 0.03% for non-multimorbidities. This significant difference between the two proportions (P-value = 4.4e−18, Figure 5A) indicates that the selected phenotypes are related to multimorbidities rather than double diseases. Moreover, this proportion for novel multimorbidities was also significantly higher than that for non-multimorbidities (P-value = 4.3e−4; Figure 5A), suggesting that our selected phenotypes contribute to the prediction of novel multimorbidities.

The differential phenotypes can help us better understand multimorbidities. For example, 'E66-Obesity' was identified to be multimorbid with 'G30-Alzheimer’s disease’ by MorbidGCN. Patients with the multimorbidity of 'E66-Obesity' and 'G30-Alzheimer’s disease’ were found to have significantly higher cystatin C (field 30720), urea (field 30670), glycated haemoglobin (HbA1c) (field 30750) in their blood, and spend more time to watch television (field 1070), compared with patients only with 'E66-Obesity’ or patients only with ‘G30-Alzheimer’s disease’ (Supplementary Table S8, see Supplementary Data available online at http://bib.oxfordjournals.org/). These results suggest that obese patients should reduce their time on watching television, which might be helpful to decrease the risk of Alzheimer’s disease. Besides, the cystatin C, urea and glycated HbA1c might serve as the blood biomarkers for future diagnosis and considered in the treatment of the multimorbidity. In addition, we observed several general risk phenotypes, especially phenotypes in the blood, household, physical...
Figure 3. Performance comparison of the population phenotype-based disease similarity with other four disease similarity calculation methods for multimorbidity prediction. Comparison results with sAB [10] (A), GeneNetRR [11] (B), HSDN [12] (C) and FusedSim [14] (D) methods on both the UKB and HuDiNe data sets. The sAB, GeneNetRR, HSDN and FusedSim methods calculated disease similarities based on the protein interaction network, reconstructed gene network, clinical symptoms and multi-omics data fusion (i.e. ontology, clinical symptom, literature occurrence, genetic association, gene expression and drug data), respectively. The top and bottom in (A), (B), (C) and (D) are the AUC and AUPRC results, respectively. Comparison results with sAB [10], GeneNetRR [11], HSDN [12] and FusedSim [14] methods for the prediction of the same and cross-physiological multimorbidities on the UKB (E) and HuDiNe (F) data sets.

activity, physical measures and sleep groups, which were differentially distributed among many types of multimorbidity patients compared with single-disease patients (Supplementary Figure S4, see Supplementary Data available online at http://bib.oxfordjournals.org/). For example, the two most common phenotypes, i.e. ‘924-Usual walking pace’ and ‘30720-Cystatin C’, were differentially distributed between multimorbidity and single-disease patients for 7172 and 4752 multimorbid disease pairs, respectively (Supplementary Results, see Supplementary Data available online at http://bib.oxfordjournals.org/, for details about the two phenotypes).

As we observed that blood phenotypes were differentially distributed for many types of multimorbidity, we analyzed whether the differential blood biomarkers (total 30 phenotypes; Supplementary Table S8, see Supplementary Data available online at http://bib.oxfordjournals.org/) were able to predict future multimorbidity diseases for patients with one disease of the multimorbid disease pairs. This analysis was based on the Cox proportional hazard model (Supplementary Methods, see Supplementary Data available online at http://bib.oxfordjournals.org/). We found a total of 3002 multimorbid disease pairs which could be predicted by these differential blood phenotypes (Supplementary Table S9, see Supplementary Data available online at http://bib.oxfordjournals.org/). Here, we took Alzheimer’s disease, a common progressive neurodegenerative disorder, to exemplify the prognostic effect of blood phenotypes
Figure 4. Performance comparison of MorbidGCN with other baseline methods and novel multimorbidities identified by MorbidGCN. Comparison results with the DeepWalk [21], Node2Vec [22], MorbidGCN−, MorbidGCN+ and MorbidPhe models for multimorbidity prediction on the UKB (A) and HuDiNe (B) data sets. MorbidGCN− and MorbidGCN+ are two variant GCN models, where the former one only uses the disease network topology information while the latter one integrates the disease network topology with all phenotypes. The MorbidPhe model only uses the selected phenotypes. The DeepWalk and Node2Vec models are two other network representation learning methods, which only use the disease network topology information for multimorbidity prediction. (C) Multimorbidity overlaps among the UKB known multimorbidities, UKB novel multimorbidities, HuDiNe known multimorbidities and HuDiNe novel multimorbidities.

Figure 5. Differential phenotypes between multimorbidity patients and single-disease patients in the UKB data set. (A) Proportions of the known multimorbidities, novel multimorbidities and non-multimorbidities, whose patients were significantly different from the single-disease patients at the selected phenotypes. The P represents the P-value generated by the two-sided t-test. (B) Prognosis of Alzheimer’s disease for patients with ’I10-Essential (primary) hypertension’ (left), ’H26-Other cataract’ (middle) or ’N39-Other disorders of urinary system’ (right) based on differential blood phenotypes using the Cox proportional hazard model. The age and sex were used as the covariates. * represents that the variable is significant for the prognosis of Alzheimer’s disease.

on multimorbidities. We found that the phenotype ‘30750-Glycated HbA1c’ could predict the occurrence of Alzheimer’s disease for patients with ’I10-Essential (primary) hypertension’ (P-value = 4.4e−09), ’H26-Other cataract’ (P-value = 2.5e−09) or ’N39-Other disorders of urinary system’ (P-value = 2.3e−08) (Figure 5B). Considering HbA1c is a diagnostic biomarker for diabetes [40], we further examined whether the prognostic effect of HbA1c on Alzheimer’s disease is caused by the diagnosis of diabetes. Through adding the diabetes status as a covariate in the Cox model, we observed that HbA1c still had the prognostic effect to Alzheimer’s disease for patients with I10 (P-value = 0.02), H26 (P-value = 0.002) or N39 (P-value = 0.003) (Supplementary Figure S5, see
The selected phenotypes tended to be differentially distributed between multimorbidity and single-disease patients. These findings have validated the effectiveness of the population phenotypes and disease network on multimorbidity prediction, while the identified novel multimorbidities and differential phenotypes can be instructive to the diagnosis, prevention and treatment of multimorbidities.

Discussion

We have developed a novel multimorbidity prediction model named MorbidGCN and demonstrated its effectiveness on two large-scale multimorbidity data sets. Different from previous methods which are based on biomolecules or literature-mined terms [10–12, 14], our model is more advantageous in two aspects: (i) the disease-phenotype associations are derived from the population cohort rather than from literature, or database-mining, which can cover more diseases and (ii) our model can effectively utilize the disease network topology, which has not been considered previously. The two types of information are integrated by GCN, resulting in a significantly improved prediction performance compared with other methods. With MorbidGCN, tens of thousands of novel multimorbidities have been identified, which can supplement the known multimorbidities derived from EHRs.

MorbidGCN treats the multimorbidity prediction as a missing link prediction task in the disease network, with the population phenotypes being diseases’ features. Owing to the high dimension of population phenotypes (>2000), a feature selection method is required to select useful phenotypes before integrating them with the disease network. However, only a few studies have performed feature selections in link prediction tasks previously [41, 42], while they are of high time complexity, as they need to quantify the associations between any two features and iteratively evaluate the feature subset. Therefore, they are not applicable to our problem. Considering the above, we proposed a feature selection method with low computation complexity. With our proposed method, the selected phenotypes can better predict the multimorbidities than all phenotypes (Figures 2E, F and 4A, B).

Population phenotypes are effective in predicting both the same and cross-physiological multimorbidities, while other biological data such as biomolecules and literature-mined terms [10–12, 14] are only effective in predicting the same physiological multimorbidities due to the incomplete knowledge on disease mechanisms (Figure 3E and F). Moreover, most of the novel multimorbidities identified by MorbidGCN are cross-physiological (Supplementary Tables S6 and S7, see Supplementary Data available online at http://bib.oxfordjournals.org/).

These results can be instructive for future studies on multimorbidities, as current healthcare systems are usually organized by focusing on specific conditions or body systems [43].

Although the GCN has achieved a remarkable success in a variety of tasks [23, 24, 26], it has limitations in learning representations of nodes with small degrees in networks [44, 45]. Consistent with this, we have found that most of the mis-predicted multimorbidities include diseases with small degrees. Therefore, our future work will be focused on designing a GCN model with a degree-aware structure to improve the prediction performance for diseases with relatively fewer multimorbidities.

Altogether, MorbidGCN integrates population phenotypes and disease network based on GCN for multimorbidity prediction and has achieved a higher prediction performance than other methods. With MorbidGCN, many novel multimorbidities have been identified. Moreover, many phenotypes have been identified to be differentially distributed between multimorbidity and single-disease patients. These findings have validated the effectiveness of the population phenotypes and disease network on multimorbidity prediction, while the identified novel multimorbidities and differential phenotypes can be instructive to the diagnosis, prevention and treatment of multimorbidities.

Supplementary Data available online at http://bib.oxfordjournals.org/.

Key Points

- We developed a multimorbidity prediction model named MorbidGCN, which integrated population phenotypes and disease network based on graph convolutional network for multimorbidity prediction. MorbidGCN outperformed all other competitive methods and was the first study to predict multimorbidities using population phenotypes.
- With MorbidGCN, many novel multimorbidities were discovered, which were highly reliable and can supplement the known multimorbidities derived from electronic health records.
- A novel feature selection method was embedded in MorbidGCN to select useful phenotypes for multimorbidity prediction.
- The selected phenotypes tended to be differentially distributed between multimorbidity patients and single-disease patients and could help interpret multimorbidities as well as showed potential for the prognosis of multimorbidities.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgment

We thank the Greater Bay Area Institute of Precision Medicine (Guangzhou).

Funding

National Key Research and Development Program of China (2020YFA0712403); National Natural Science Foundation of China (61932008); Shanghai Municipal Science Technology Major Project (2018SHZDZX01); 111 Project (No. B18015) of China.

Code availability

The source code of this work can be downloaded from GitHub (https://github.com/ZhaoXM-Lab/MorbidGCN).
Data availability
The health and phenotype data used in this study are available from the UKB with restrictions applied. Data were used under license and thus not publicly available. Access to the UKB data can be requested through a standard protocol (https://www.ukbiobank.ac.uk/register-apply/). Other data can be accessed through the following sites: the summary statistics of disease pairs in the human disease network are from http://sbi.upf.edu/data/hudine/; the separation ($s_{AB}$) of disease pairs in the human interactome is from https://www.science.org/doi/suppl/10.1126/science.1257601/suppl_file/datasets_s1-s4.zip; the disease embeddings based on the gene network reconstruction and representation (GeneNetRR) were generated by the codes and data provided in https://github.com/catly/disease_similarity; the term co-occurrences between symptoms and diseases in human symptoms-disease network (HSDN) are from https://static-content.springer.com/esm/art%3A10.1038%2Fncommms5212/MediaObjects/41467_2014_BFncommms5212_MOESM1045_ESM.txt; the fused similarity scores between disease pairs (FusedSim) based on multi-omics data fusion were generated by the codes and data provided in https://github.com/e-oerton/disease-similarity-fusion.

References


