Computational Algorithms for Multi-omics and Electronic Health Records Data

Jia Guo

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Computational Algorithms for Multi-omics and Electronic Health Records Data

Jia Guo

Real world data have enhanced healthcare research, improving our understanding of disease progression, aiding in diagnosis, and enabling the development of personalized and targeted treatments. In recent years, multi-omics data and electronic health record (EHR) data have become increasingly available, providing researchers with a wealth of information to analyze. The use of machine learning methods with EHR and multi-omics data has emerged as a promising approach to extract valuable insights from these complex data sources. This dissertation focuses on the development of supervised and unsupervised learning methods, as well as their applications to EHR and multi-omics data, with a particular emphasis on early detection of clinical outcomes and identification of novel cancer subtypes.

The first part of the dissertation centers on developing a risk prediction tool using EHR data that enables disease early detection so that preventive treatments can be taken to better manage the disease. For this goal, we developed a similarity-based supervised learning method with two applications to predict end-stage kidney disease (ESKD) and aortic stenosis (AS). In the second part of the dissertation, we expanded our goal to a phenome-wide prediction task and developed a patient representation based deep learning method that is able to predict phenotypes across the phenome. Through a weighting scheme, this approach is conducting tailored disease phenotype prediction computationally efficiently with good prediction performance. In the final part of the dissertation, I shifted the focus with the goal to identify clinical meaningful novel disease subtypes

with unsupervised learning methods using multi-omics data. We tackled this goal through integrating multiple patient graphs being generated from multiple omics data with molecular level features for an improved disease subtyping.

This dissertation has significantly contributed to the development of data-driven approaches to healthcare and biomedical research using EHR data and multi-omics data. The new methodologies developed with applications in multiple diseases using EHR and multi-omics data advanced our knowledge in disease diagnosis, vulnerable groups identification, and ultimately improve patient care.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I wish to express my deepest gratitude to those who have steadfastly supported me throughout the journey of completing this dissertation. Their assistance, guidance, and encouragement have been invaluable in the development of my dissertation.

First and foremost, I would like to thank my PhD advisor, Dr. Shuang Wang, for her unwavering support and mentorship. Her profound expertise, insightful advice, and remarkable patience were crucial in shaping my research and fostering my intellectual growth. Without her wisdom and guidance, this dissertation would not be possible.

I extend my sincere appreciation to my dissertation committee members, Dr. Chunhua Weng, Dr. Krzysztof Kiryluk, Dr. Yuanjia Wang, and Dr. Min Qian. Their time, support, and feedback have been of immense value. Dr. Chunhua Weng provided essential assistance and expertise on one chapter of my thesis. Dr. Krzysztof Kiryluk provided constructive comments and advice on two chapters. Their invaluable input and contribution significantly elevated the quality of my work. Dr. Yuanjia Wang and Dr. Min Qian not only imparted knowledge to me in the courses, but also provided insightful feedback and suggestions that greatly improved my dissertation. Their mentorship and guidance played a critical role in ensuring the success of my research.

I also want to thank Dr. Frederica Perera and Dr. Julie Herbstman for their financial sponsorship throughout my PhD journey. We collaborated on various environmental health projects and their support has substantially enriched my research experience and professional growth.

Lastly, I want to express my gratitude to my family. I am profoundly grateful to my parents, who continually served as my pillar of strength. My deepest appreciation also goes out to my fiancée, Yun He. Her unwavering support and encouragement have carried me through the challenging moments of my doctoral journey.

Chapter 1: Introduction

1.1 Overview

The enormous amount of healthcare data generated every day provides a unique opportunity to enhance our understanding of disease progression, aid in diagnosis, and develop more personalized and targeted treatments. Electronic health record (EHR) data together with multiple types of omics data represent a valuable resource for big data researches including clinical decision-making and disease subtyping to better understand disease heterogeneity and improve patient care through targeted treatments to early detection. By leveraging the power of machine learning and deep learning methods as well as statistical modeling techniques, big data research that can help clinicians make more informed decisions, ultimately leading to better patient care. This dissertation seeks to develop computational algorithms to build clinical decision-support tools using EHR and multi-omics data.

We first focused on risk prediction of specific clinical outcomes using EHR data. We introduced a framework named Patient similarity based on Domain Fusion (PsDF). Using comprehensive patient data, PsDF integrates similarity information from multiple clinical data domains into a comprehensive similarity measurements that can be subsequently used to predict clinical outcomes. PsDF is a supervised machine learning method, aiming to aid in early detection of specific diseases. We used PsDF to predict end stage kidney disease (ESKD) and severe aortic stenosis (AS) requiring valve replacement. We demonstrated good prediction performance of PsDF and robustness of PsDF to missingness compared to comparison methods using the data warehouse of Columbia University Irving Medical Center (CUIMC).

We then expanded our goal to phenome-wide predictions, to predict phenotypes in a phenome simultaneously. We developed a Phenome-wide prediction framework using Weighted Patient

Vectors (PheW^{2}P2V). PheW^{2}P2V uses a novel weighting scheme on patient embeddings so that predictions based on patient embeddings are tailored to individual phenotypes. Since diagnosis concepts in EHR are usually coded using International Classification of Disease (ICD) terminology, which is designed for billing and administrative functions but not for case-control studies, PheW^{2}P2V first maps patients' ICD codes to clinical disease phenotypes called phenotype codes (phecodes). Phecodes are originally developed for phenome-wide association studies (PheWAS), where patients' phenotypes are identified by grouping ICD codes that represent common etiologies, with a purpose of reducing the redundancy in ICD codes and better defining cases and controls. To predict a clinical disease phenotype in the phenome, after generating medical concepts embeddings using word2vec, PheW^{2}P2V represents each patient as a single numeric patient vector where his(her) medical concepts that are more correlated with the phenotype of interest will be upweighted. The tailored patient vector is then used to predict risks. We demonstrated improved predictions of PheW^{2}P2V over that of comparison methods for majority of the phenotypes in the phenome of the MIMIC-III database (Johnson et al., 2016) performing of 942 phenotypes.

In the third part of the dissertation, we aim to develop computational algorithm to identify disease subtypes using multi-omics data. We developed a multi-view clustering method for disease subtyping, Multi-view Graph Convolutional Clustering (MultiGCC). Many existing disease subtyping methods leverage patient graphs that are generated using similarity measures from high dimensional multi-omics data. These patient graphs focus on subject level aggregated omics information but ignore feature level individual molecular characteristics, which are also helpful for disease subtyping but are absent in graphs. MultiGCC uses feature level molecular characteristics to enhance patient omics graph embeddings through graph convolutional encoders. MultiGCC then simultaneously updates the graph embeddings and the clustering assignments through a self-learning process to achieves a better separation of clusters. We applied MultiGCC to use somatic mutation, DNA methylation, and gene expression data to identify subtypes of liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD) from The Cancer Genome Atlas (TCGA) project (https://www.cancer.gov/tcga). Further analyses using molecular characteristics suggested

clinical meaning of the identified subtypes of LIHC and STAD.

In summary, in this dissertation, we developed i) a similarity-based supervised learning method for risk prediction of specific diseases using EHR data, ii) an embedding-based supervised learning method for phenome-wide predictions using EHR data, and iii) a multi-view unsupervised learning method for disease subtyping using integrated multi-omics data. The three new methodologies developed with applications in multiple diseases using EHR and multi-omics data advanced our understanding in disease heterogeneity, diagnosis, and ultimately help to improve patient care.

1.2 EHR and multi-omics data

Electronic health record (EHR) data refers to the digital, systematic collection of patients' medical history, diagnoses, medications, treatment plans, and other relevant health information, which are maintained and updated by healthcare providers across multiple care settings (Agrawal & Prabakaran, 2020; Dash et al., 2019; Raghupathi & Raghupathi, 2014; Shivade et al., 2014). EHR system aims to improve the overall quality, safety, and efficiency of healthcare delivery by providing clinicians with real-time access to comprehensive patient information, facilitating informed decision-making, and reducing medical errors (Menachemi & Collum, 2011). The adoption of EHR systems has increased rapidly over the past two decades, with recent estimates suggesting that approximately 88% of office-based physicians in the United States utilize EHR data (Office of the National Coordinator for Health Information Technology, 2021).

Multi-omics data refers to the various layers of high-dimensional biological information generated through different '-omics' technologies, including genomics, transcriptomics, proteomics, and metabolomics (Hasin et al., 2017; Kristensen et al., 2014; Rappoport & Shamir, 2018). Each of these data types captures specific molecular features of an organism or biological system: genomics focuses on gene sequences and variations; transcriptomics investigates RNA transcripts and gene expression profiles; proteomics examines protein abundance and post-translational modifications; and metabolomics explores the concentrations of metabolites in biological pathways (Aebersold & Mann, 2016; Emwas et al., 2019). By using multi-omics data, researchers can uncover complex

relationships and interactions between various biological components, which facilitates a deeper understanding of the underlying mechanisms of health and disease (Subramanian et al., 2020). For example, the integration of multi-omics data with advanced computational algorithms has opened up new avenues for better understanding of tumor heterogeneity and improving personalized treatments through cancer subtyping (B. Wang, Mezlini, et al., 2014).

1.3 Supervised learning methods for classification

Supervised learning is a sub-field of machine learning that focuses on the development of algorithms capable of learning from labeled data, and subsequently using this learned knowledge to make predictions or decisions. Supervised learning has been extensively employed in a wide range of applications, from natural language processing (NLP), computer vision, to healthcare and finance (Bishop & Nasrabadi, 2006; Kelleher et al., 2020). A supervised learning model is usually described as a mathematical representation of the underlying relationships between input features and the target variable. In specific, supervised learning methods try to find a function f_θ with parameter θ to map input features to a target variable based on a given set of labeled training data pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is the input feature vector and y is the target. Suppose we have n training samples $\{(\boldsymbol{x_1}, y_2), ..., (\boldsymbol{x_n}, y_n)\}$, then the goal is to learn a function $f_\theta : X \rightarrow Y$ that best approximates the true underlying relationship between $\boldsymbol{x_i} \in X$ and $y_i \in Y$, such that for any new input instance $\boldsymbol{x'}$, we can make an accurate prediction $y' = f_\theta(\boldsymbol{x'})$. To achieve this, the model is trained by minimizing a loss function $\mathcal{L}$, which quantifies the discrepancy between the model predictions $f_\theta(\boldsymbol{x_i})$ and the true target values y_i in the training dataset, showed in Equation 1.1.

$$\underset{f_\theta}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f_\theta(\boldsymbol{x_i})) \tag{1.1}$$

The loss function $\mathcal{L}$ depends on the type of supervised learning problem. There are two fundamental types of problems, regression and classification. Regression deals with predicting a continuous target variable such as BMI and blood pressure, while classification focuses on assigning

discrete class labels to input instances. In this dissertation, we mainly deal with binary clinical outcomes such as case-control status of patients regarding a specific disease, hence, we focus on the supervised learning methods for classification. Supervised learning methods for classification can be broadly categorized into probabilistic models, kernel-based models, tree-based models, and artificial neural networks.

Naive Bayes and logistic regression are two commonly used probabilistic models. Naive Bayes focuses on modeling the joint distribution of the input features and the target variables, by applying Bayes' theorem with naive independence assumptions between the features. Although it is highly scalable, the design and assumptions are apparently oversimplified. Logistic regression estimates the conditional probabilities of the target variables given the input features and makes predictions based on these probabilities. It employs the logistic function (sigmoid function σ) to map a linear combination of input features $\boldsymbol{\beta}^\top \boldsymbol{x_i}$ to a probability value p_i with model parameters $\boldsymbol{\beta}$, and minimizes the cross-entropy loss function (Bishop & Nasrabadi, 2006), showed in Equation 1.2. The optimization is also equivalent to the maximum likelihood estimation. Both Naive Bayes and logistic regression lead to a linear decision boundary, hence, they are less powerful when deal with non-linearly separable data.

$$
\begin{aligned}
p_i(y_i = 1|x_i) = \sigma(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x_i}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x_i})}} \\
\mathcal{L}(y_i, p_i) = -\left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right]
\end{aligned}
\tag{1.2}
$$

Kernel-based models, or similarity-based methods in a broader sense, are popular in classification tasks, as they can handle non-linearly separable data. For example, the k-nearest neighbors (kNN) is a similarity-based, lazy learning algorithm that predicts the target label for a new sample by considering the k closest training samples using similarity measures (Cover & Hart, 1967). Support vector machines (SVMs) are powerful classifiers that can find the optimal hyperplane separating different classes in high-dimensional spaces by leveraging kernel functions to perform a non-linear classification (Cortes & Vapnik, 1995). The hinge loss function (Equation 1.3) is

usually used in SVMs.

$$\mathcal{L}(y_i, f_\theta(\boldsymbol{x_i})) = \max(0, 1 - y_i f_\theta(\boldsymbol{x_i})) \tag{1.3}$$

Tree-based models are also widely used because they are highly intuitive and can construct complex non-linear decision boundaries. The basic model is the decision tree, which recursively partitions the input feature space into regions using a series of simple decision rules (e.g., whether a feature value is greater than a threshold) at each internal node of a tree (Quinlan, 1986). It employs loss functions that assess the split based on the purity of the resulting nodes, such as Gini impurity (Breiman, 1984) and information gain (Quinlan, 1986). One disadvantage of decision tree is that it is prone to overfitting, and ensemble learning can help avoid overfitting. Random forest is an ensemble method that construct multiple decision trees and combine their predictions via majority voting in classification tasks. The trees are constructed using bootstrapped samples of the training data and a random subset of features, reducing the impact of overfitting and improving overall model performance (Breiman, 2001). Gradient boosting tree is another ensemble method that combines a series of shallow decision trees in a sequential manner. Each subsequent tree is trained to correct the errors of the previous one by fitting the negative gradient of the loss function (Friedman, 2001).

The last category is Artificial Neural Networks (ANNs). Inspired by the biological nervous system, ANNs are capable of learning complex non-linear relationships through the composition of interconnected nodes or neurons. Deep neural networks also refers to the deep learning models, as they consist of multiple processing layers to learn representations of data with multiple levels of abstraction, with each neuron applying a non-linear activation function to the weighted sum of its inputs. (Goodfellow et al., 2016; LeCun et al., 2015). The most common optimization technique for training deep learning models is the backpropagation algorithm, which minimizes the loss function using gradient descent (Rumelhart et al., 1986). Equation 1.4 is the general formula of a deep learning model with n layers for classification, where $\{W_1, W_2, ...\}$ are the weight matrix (model parameters) to be estimated, $\{f_1, f_2, ...\}$ are predefined non-linear functions such as sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ and rectified linear unit (ReLU) functions $\text{ReLU}(x) =$

$\max(0, x)$, and intercept terms are omitted for simplicity.

$$\begin{aligned} \hat{y}_i &= f_n(...f_2(W_2 f_1(W_1 \boldsymbol{x_i}))) \\ \underset{\{W_1, W_2, ...\}}{\text{minimize}} & \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) \end{aligned} \tag{1.4}$$

Deep learning methods have been applied in the field of classification, achieving state-of-the-art performance on a wide range of tasks. One of the most prominent deep learning models for classification is the Convolutional Neural Network (CNN), which has been particularly successful in image classification tasks (Krizhevsky et al., 2017). CNNs incorporate convolutional layers, pooling layers, and fully connected layers to learn spatial hierarchies of features in a translation-invariant manner. Another notable deep learning model is the Recurrent Neural Networks (RNNs) and its variants, such as Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014). These models have shown remarkable performance in sequence classification tasks, including natural language processing (NLP) and time-series analysis, by effectively modeling long-range dependencies in the data. More recently, the Transformer architecture (Vaswani et al., 2017) has gained significant attention due to its superior performance on various classification tasks, particularly in natural language understanding. The Transformer model relies on self-attention mechanisms to capture complex relationships within the input data, making it a powerful and versatile tool for various classification problems. These deep learning methods have demonstrated their efficacy across numerous domains and continue to advance the state-of-the-art in classification tasks.

In chapter 2, we developed a similarity-based supervised learning methods (PsDF) using EHR data to predict two diseases, ESKD and AS. The proposed PsDF leverages the advantage of similarity-based methods that can effectively capture the non-linear signals and uses a fusion step to achieve a comprehensive similarity measure. In chapter 3, we developed an embedding-based supervised learning method (PheW^{2}P2V) for phenome-wide predictions using EHR data. The proposed PheW^{2}P2V generates patient embeddings using a deep learning model and efficiently

performs phenotype predictions across the phenome.

1.4 Unsupervised learning methods for clustering

Unsupervised learning is another sub-field of machine learning that has emerged as a powerful tool for discovering hidden patterns and structures within data but without relying on labeled instances for guidance. There are two important types of tasks for unsupervised learning, clustering and dimension reduction. Clustering is used to group or segment data points into distinct categories with shared attributes, while dimension reduction is used to transform the data from a high-dimensional space into a low-dimensional space. In modern machine learning, the dimension reduction techniques can also be considered as a special type of representation learning, which aims to learn new representations of data and improve the performance and interpretability of downstream analysis such as classification, clustering, and data visualization. In this dissertation, we mainly consider clustering tasks such as cancer subtyping using high-dimensional omics data, hence, we focus on the unsupervised learning methods for clustering with the help of representation learning.

Unsupervised learning methods for clustering can be broadly categorized into partition-based clustering, probabilistic clustering, density-based clustering, hierarchical clustering, and a combination of representation learning and partition-based clustering. Partition-based clustering methods, such as K-means (MacQueen, 1967) and its variants Fuzzy C-means (FCM) (Bezdek et al., 1984) as well as K-means++ (Arthur & Vassilvitskii, 2006), focus on dividing the data into distinct groups or partitions based on a distance metric. These partition-based algorithms typically aim to minimize the within-cluster sum of squares, iteratively adjusting the cluster centroids until convergence. Probabilistic clustering is a family of algorithms that assume data points are generated from a mixture of underlying probability distributions, such as Gaussian Mixture Models (GMM) (Dempster et al., 1977), Bayesian Gaussian Mixture Models (BGMM) (Richardson & Green, 1997) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Density-based clustering algorithms, like DBSCAN (Ester et al., 1996) and OPTICS (Ankerst et al., 1999), identify clus-

ters based on the density of data points in the feature space, where they group data points that are closely packed together while treating noise points or outliers as separate entities. Hierarchical clustering methods (Murtagh & Contreras, 2012), such as agglomerative and divisive hierarchical clustering, build a tree-like structure to represent the nested grouping of data points.

The last category gain more attention recently, which includes clustering methods that combine representation learning and partition-based clustering. These methods usually learn a low dimensional latent space or a new representation of the data and perform the standard K-means using the latent representations. The representation learning algorithms could be performed either on the raw feature space, such as principal component analysis (PCA), non-negative matrix factorization (NMF) (Lee & Seung, 1999), and Autoencoder (Bengio et al., 2013; Hinton & Salakhutdinov, 2006; Vincent et al., 2010), or on a similarity matrix constructed from the raw input, such as spectral clustering (i.e., eigen decomposition on the graph) (Von Luxburg, 2007) and Autoencoder performed on a graph (F. Tian et al., 2014).

As a popular method in this category, spectral clustering partitions data into clusters by leveraging the eigenvectors derived from a graph and performing the standard K-means. In biomedical research, a graph is usually characterized by a similarity matrix measuring the affinity between a pair of instances, e.g., a patient similarity matrix constructed using omics data. The purpose of using similarity measures is to model the local neighborhood relationships among data points and avoid pre-screening feature selection on the high-dimensional feature space of raw data (Von Luxburg, 2007; B. Wang, Mezlini, et al., 2014). Suppose we have a similarity matrix $S \in \mathbb{R}^{n \times n}$ and k clusters to construct, then we first calculate a normalized Laplacian L using Equation 1.5, where D is the degree matrix with $D_{ii} = \sum_j S_{ij}$ and I is the identity matrix. Next, first k eigenvectors of L are computed and formed $Z \in \mathbb{R}^{n \times k}$, which can be considered as the new representations (spectral embeddings) of the data. K-means algorithm is then performed on Z to cluster the data points (Ng et al., 2001; Von Luxburg, 2007).

$$L = I - D^{-1/2} S D^{-1/2} \tag{1.5}$$

Autoencoder combined with K-means is also widely used for clustering (Fang et al., 2021; X. Guo et al., 2017; Hinton & Salakhutdinov, 2006; Miotto et al., 2016; Salha et al., 2019; F. Tian et al., 2014). Autoencoder is a powerful representation learning method, which uses a neural network to learn new representations of data from raw features, typically for dimension reduction. It uses a non-linear function mapping (encoder) to map the raw features to new representations and another non-linear function mapping (decoder) to reconstruct the raw features. The parameters of the model can be trained by minimizing the reconstruction error. When using affine encoder and decoder without any non-linearity and a squared error loss, the Autoencoder essentially performs PCA, where the trained weights span the same subspace as the one spanned by the loading vectors of PCA (Plaut, 2018). Equation 1.6 is the general formula of an Autoencoder with one hidden layer, where $\boldsymbol{x_i}$ and $\hat{\boldsymbol{x_i}}$ are the raw input and reconstructed output, $\boldsymbol{z_i}$ is the new representations with a lower dimension than $\boldsymbol{x_i}$, f and g are non-linear functions (e.g., sigmoid and ReLU function) for encoder and decoder, W_e and W_d are the weight matrix to be learned by minimizing the reconstruction loss function $\mathcal{L}$ with gradient descent. The new representation $\boldsymbol{z_i}$ can be subsequently used to perform clustering through K-means. A stacked Autoencoder is an Autoencoder with more than one hidden layer, which can increase the nonlinearity between the new representations and raw features. For example, in a stacked Autoencoder with a two-layer encoder, the new representations will be calculated as $\boldsymbol{z_i} = f_2(W_{e2} f_1(W_{e1}\boldsymbol{x_i}))$.

$$
\begin{aligned}
\boldsymbol{z_i} &= f(W_e \boldsymbol{x_i}) \\
\hat{\boldsymbol{x_i}} &= g(W_d \boldsymbol{z_i}) \\
\underset{\{W_e, W_d\}}{\text{minimize}} \; & \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{x_i}, \hat{\boldsymbol{x_i}})
\end{aligned}
\tag{1.6}
$$

In the previous methods, learning new representations and conducting clustering are usually performed sequentially. That is, K-means clustering is applied after a new representation of data is obtained by Autoencoder or spectral embedding. Although it is possible to produce a better clustering result using the new representations than using the raw data, there is still room for im-

provement. Simply minimizing the reconstruction loss may not guarantee different clusters to be linearly separable in the space of new representations, while K-means is incapable of separating clusters that are not linearly separable. Deep Embedded Clustering (DEC) has been proposed to optimize a clustering objective by simultaneously solving for cluster assignment and the underlying feature representation (Xie et al., 2016). DEC improves the clustering by iteratively updating clusters with an auxiliary target distribution. Inspired by t-SNE (Van der Maaten & Hinton, 2008), the loss function of DEC is a Kullback–Leibler (KL) divergence loss between a centroid-based probability distribution and its auxiliary target distribution. Equation 1.7 is the formula of DEC, where i and r represent the ith sample and the rth cluster, respectively. Let $\boldsymbol{z_i}$ be a new representation of data obtained from a representation learning method such as Autoencoder. With K-means applied on the new representations, the initial cluster centroid of rth cluster can be estimated as $\boldsymbol{\mu_r}$. The centroid-based probability distribution Q, or the soft cluster assignments q_{ir} is calculated using the kernel of Student's t-distribution with degrees of freedom α (usually set as $\alpha = 1$), to measure the similarity between $\boldsymbol{z_i}$ and $\boldsymbol{\mu_r}$. To put more emphasis on data points with high confidence or high q_{ir}, the auxiliary target distribution p_{ir} is computed by first raising q_{ir} to the second power and then normalizing by frequency for each cluster, where $v_r = \sum_i q_{ir}$ are soft cluster frequency of the rth cluster.

$$
\begin{aligned}
q_{ir} &= \frac{\left(1 + \|\boldsymbol{z_i} - \boldsymbol{\mu_r}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{r'} \left(1 + \|\boldsymbol{z_i} - \boldsymbol{\mu_{r'}}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \\
p_{ir} &= \frac{q_{ir}^2 / v_r}{\sum_{r'} q_{ir}^2 / v_{r'}} \\
&\text{minimize } \mathrm{KL}(P\|Q)
\end{aligned}
\tag{1.7}
$$

In Equation 1.7, all model parameters including the cluster centroids $\boldsymbol{\mu_r}$ and weight matrix of the encoder that generates $\boldsymbol{z_i}$ will be updated by minimizing the KL divergence loss between Q and P, where $\mathrm{KL}(P\|Q) = \sum_i \sum_r p_{ir} \log\left(\frac{p_{ir}}{q_{ir}}\right)$. The target distribution P is designed so that the training step will put more emphasis on data points that are assigned with higher confidence, or higher q_{ir}.

By updating model parameters from high confidence assignments, DEC iteratively refines the soft cluster assignments q_{ir}, which is the final clustering result.

In chapter 4, we developed a multi-view unsupervised learning method (MultiGCC) for disease subtyping using multi-omics data. The proposed MultiGCC generates enhanced graph embeddings of omics data by incorporating molecular level features. MultiGCC can identify novel disease subtypes by simultaneously updating the graph embeddings and clustering assignments through a self-learning process.

Chapter 2: Similarity-based health risk prediction using domain fusion and electronic health records data

2.1 Introduction

The universal adoption of electronic health records (EHR) provides access to clinical data of unprecedented volume and variety. This rich information awaits utilization for real time clinical decision-support. Conventional approaches in predictive modeling used to build clinical decision-support tools start with feature selection based on domain knowledge, which could be biased. For example, one of the most widely used chronic kidney disease (CKD) progression models uses a simple linear combination of age, sex, estimated glomerular filtration rate (eGFR), and urinary albumin to creatinine ratio (UACR) (Tangri et al., 2016). This CKD prediction model and other similar existing prediction models were built on a clinically relevant set of features selected either based on clinical expertise, statistical significance, or both. As evidenced from recent scientific research, many human disorders share a complex etiologic basis and exhibit correlated disease progression. Therefore, it is desirable to consider a more comprehensive, agnostic approach that incorporates the entirety of patient data.

One frequently sought goal by using EHR data is to assess patient similarity (Chan et al., 2010; Chawla & Davis, 2013; Jensen et al., 2012; L. Li et al., 2015; Marlin et al., 2012; Miotto et al., 2016; Roque et al., 2011; Sun et al., 2012; F. Wang, Hu, et al., 2012; F. Wang, Sun, & Ebadollahi, 2011, 2012; F. Wang, Sun, Hu, et al., 2011; P. Zhang et al., 2014; Zhu et al., 2016). The objective of patient similarity assessment is to quantify the similarity between any pair of patients according to their retrospective information under a specific clinical context. For example, patients who have "similar" clinical characteristics may have similar disease risk projections or diagnoses. Similarity-based case identification could help stratify patients, enable more efficient diagnoses, and facilitate

more effective treatment choices. Despite some successes, current similarity approaches do not use comprehensive patient information, but rather only a fraction of available data, such as only selected clinical characteristics or only genomic information for patient subtyping (L. Li et al., 2015). A recent approach combined clinical and drug similarity analytics for personalization of drug prescribing (P. Zhang et al., 2014). Another recent research developed a disease phenotyping method with tensor factorization using co-occurrence information of diagnoses and medications (Henderson et al., 2018). Phenotyping algorithms use EHR data to identify patients with specific clinical conditions or events. These include rule-based algorithms to identify patients with chronic kidney disease (CKD) (Nadkarni et al., 2014), supervised models including logistic regressions and random forest to identify patients with type 2 diabetes (Anderson et al., 2016), and dimensionality-reduction methods such as a tensor factorization approach to identify patients with hypertension and type 2 diabetes (Henderson et al., 2018). Several recently developed phenotyping methods also consider patient similarities, such as a pipeline that defines patient similarities using concatenated patient concepts in Unified Medical Language System (UMLS) which was applied to ciliopathies phenotyping (X. Chen et al., 2019). However, there are currently no methods that use all available patient data to more comprehensively define "similar patients" for predictive outcome modeling in chronic complex conditions.

A simple way to use comprehensive patient data is to define patient similarity using patient information concatenated. However, the patient information from different domains might be unbalanced. For example, the number of unique drugs, i.e., number of features, in the domain of drug exposures might be very different from the number of unique procedures in the domain of medical procedures. Thus, when using features from these unbalanced domains, simply concatenating all features to calculate patient similarity may be ineffective in capturing signals when these potentially much stronger signal features from a small domain might be diluted.

In this chapter, we developed a unified machine learning framework for clinical outcome prediction called Patient similarity based on Domain Fusion (PsDF). PsDF performs patient similarity assessment independently on each available domain data, such as laboratory tests, ICD based di-

agnoses, drug exposures, medical procedures, and demographic information, and fuses affinity information from all available domains to achieve a comprehensive metric for quantifying patient similarity, which is further used to perform a clinical outcome prediction.

We conducted extensive simulation studies and demonstrated a much-improved prediction performance of the PsDF algorithm over several competing methods including a random forest classifier and a regression-based model both using all features from different domains simultaneously, and a naïve similarity method concatenating all features from different domains.

With EHR data extracted from the data warehouse of Columbia University Irving Medical Center (CUIMC), we demonstrated better performance of PsDF over the competing methods in predicting two independent clinical outcomes, incident end stage kidney disease (ESKD) and incident aortic stenosis (AS) requiring valve replacement. We used comprehensive patient information collected prior to the occurrence of the ESKD and AS outcomes, including 1) laboratory tests, 2) ICD based diagnosis history, 3) drug exposures, 4) medical procedures and 5) demographic information.

Because real-life EHR datasets often have incomplete patient records, we also explored the prediction robustness of PsDF when random missingness was introduced to the test set data. To do so, we randomly masked a percentage of EHR records by setting them to missing, similar to prior studies (Z. Hu et al., 2017; Polubriaginof et al., 2018; Wells et al., 2013). Our results indicate that when the percentage of randomly masked observations increases, the prediction performance of PsDF is stable while that of the competing methods decreases fast, indicating that one of the major advantages of PsDF is its robustness to data missingness.

2.2 Methods

2.2.1 The PsDF algorithm

The PsDF framework is illustrated in Figure 2.1. There are three steps in the PsDF clinical outcome prediction. In Step 1, for each domain of patient data (e.g., laboratory tests, diagnosis history, etc.), a patient similarity matrix with pairwise similarity measures between any given pairs

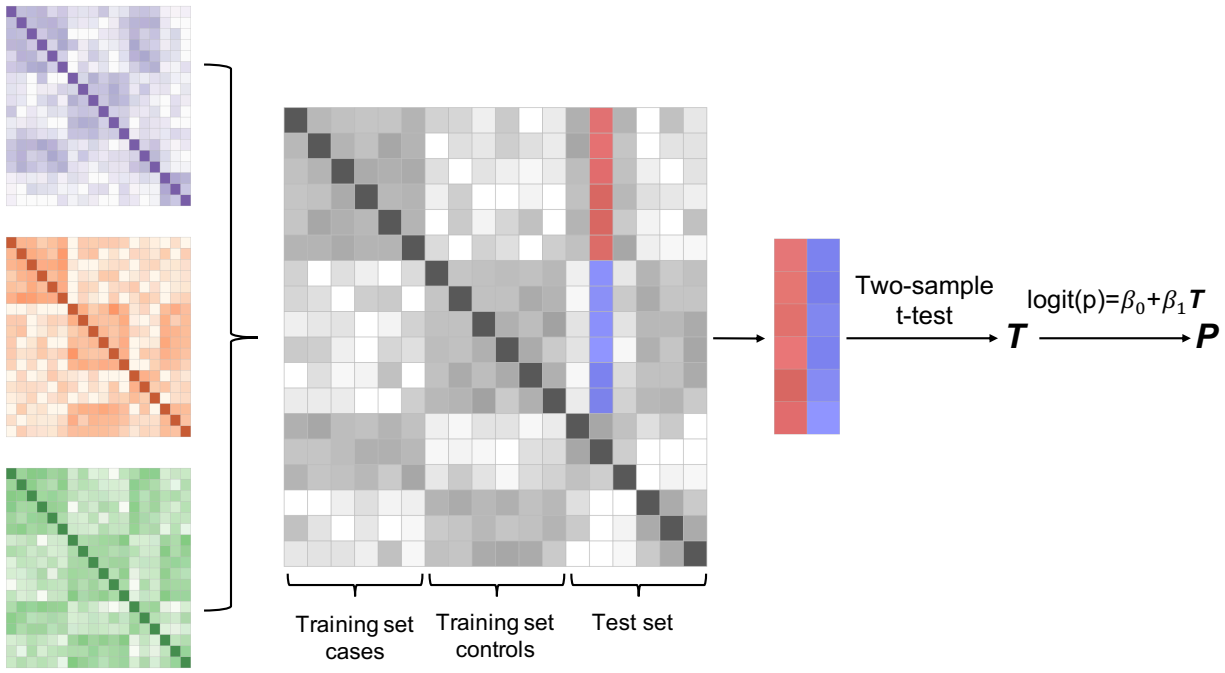

Figure 2.1: The workflow of the proposed PsDF framework.

of patients is constructed. In Step 2, patient similarity matrices from different domains of patient data are fused using a nonlinear combination method. In Step 3, the fused patient similarity matrix is served as a clinical outcome prediction tool, through which a patient similarity risk score is assigned to each patient in the test set using a simple logistic regression model that is pretrained on the training set. Note that all features of a specific domain of patient data are first standard-normalized to have a zero mean and a unit of one standard deviation.

Step 0: EHR patient's snapshot data

Patients' EHR data were extracted from the data warehouse of CUIMC. For a specific clinical condition, such as incidence of ESKD between year 2006 and 2016, in order to develop a prospective prediction model, we used a pseudo-prospective study design, where we used a snapshot of patients' retrospective EHR information from year 2006 and prior. This snapshot of EHR data includes five patient data domains: 1) laboratory tests, 2) ICD based diagnosis history, 3) drug exposures, 4) medical procedures and 5) demographic information.

We next converted EHR snapshot data (2006 and prior) into five data matrices representing

information from these five domains. Features in the four clinical domains were coded as binary features with 1 indicating a patient ever had a specific condition in 2006 and prior. Specifically, for one patient domain, e.g., drug exposures, if we have total N patients and total P possible drugs, we generated a drug exposure matrix $Y^{N\times P}$ with each row representing a patient and each column representing a drug exposure. We then considered binary status of each of the P possible drugs. For example, if there is a record in EHR data that a patient had ever taken Aspirin in 2006 and prior, and another record of patient ever taking Ibuprofen in 2006 and prior, then in the drug exposure matrix $Y^{N\times P}$, there would be one column indicating whether Aspirin had ever been taken (taken will be coded as 1) and another column indicating whether Ibuprofen had ever been taken, in 2006 and prior. We assume that a patient was not on a specific medication if there is no record in the EHR snapshot 2006 and prior. Other three patient domains, ICD based diagnosis history, laboratory tests, and medical procedures were similarly processed to generate corresponding data matrices. In the two clinical applications on incident ESKD and AS, we implemented a random mask procedure which randomly changes a certain percentage of observed records (coded as 1) to missing or unobserved (coded as 0) to explore the robustness of PsDF to missing data. Similar procedures have been applied to evaluate methods when outcomes were randomly changed to be unknown (Polubriaginof et al., 2018). For the single patient domain, demographic information has two binary variables, gender and race (coded as white or non-white).

Step 1: Constructing a patient similarity matrix for individual patient domain data

Before calculating patient similarities from the data matrix $Y^{N\times P}$, a normalization procedure is performed to normalize each column to have mean 0 and standard deviation 1. Denote $X^{N\times P}$ as the normalized matrix, for each domain of patient data, we calculate the distance between patients i and j as follows:

$$d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\| = \sqrt{\sum_{p=1}^{P}\left(x_{ip} - x_{jp}\right)^2} \tag{2.1}$$

A similarity matrix of a patient cohort with sample size N is an N by N symmetric matrix

$S^{N\times N}$, where the entry s_{ij} represents the similarity measure between patients i and j. A similarity measure quantifies the affinity between two patients. For example, a typical similarity measure can be the reciprocal of a Euclidean distance $s_{ij} = -d(\boldsymbol{x}_i, \boldsymbol{x}_j)$. It can also be a more complex measure of similarity if we use other transformation such as the radial basis function (RBF) kernel:

$$\begin{aligned} s_{ij}^{(RBF)} &= \frac{1}{\sqrt{2\pi\eta_{ij}^2}} \exp\left(\frac{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}{2\eta_{ij}^2}\right) \\ \eta_{ij} &= \frac{\mu}{3}\left[\text{mean}\left(d\left(\boldsymbol{x}_i, N_i\right)\right) + \text{mean}\left(d\left(\boldsymbol{x}_j, N_j\right)\right) + d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right], \end{aligned} \tag{2.2}$$

where μ is a hyperparameter, N_i denotes the set of nearest neighbors of patient i with a pre-fixed size of K, mean$(d(\boldsymbol{x}_i, N_i))$ is the average distance between patient i and the neighbors N_i, and η_{ij} is a scaling parameter that adapts to the density of neighbor sets so that a smaller η_{ij} is used in a denser neighbor set.

The above steps can be similarly applied to each of the individual data domain such as laboratory tests, ICD based diagnosis history, and demographic information etc., and obtain multiple patient similarity matrices. Because there are different numbers of features in different patient domains, the scales of similarity matrices S might be different. Therefore, a normalization on similarity matrices is needed. For similarity measures s_{ij} between patients i and j, we normalize as follows:

$$\begin{aligned} s_{ij} &= \begin{cases} \frac{s_{ik}^{(RBF)}}{2\sum_{k\neq i}^{N} s_{ik}^{(RBF)}}, j \neq i \\ \frac{1}{2}, j = i \end{cases} \\ s_{ij} &= \frac{1}{2}\left(s_{ij} + s_{ji}\right) \end{aligned} \tag{2.3}$$

The normalized similarity measures have a range $(0, 1)$. We denote the normalized similarity matrix as $S^{N\times N}$. These similarity matrices can be considered as similarity networks for patients whose nodes are patients and edges are similarity measures between any given pair of patients.

Step 2: Fusing patient similarity matrices from multiple patient domains data

The algorithm that fuses networks was originally developed in the field of computer vision (Blum & Mitchell, 1998; B. Wang et al., 2012; B. Wang, Mezlini, et al., 2014). For -omics

research, the Similarity Network Fusion (SNF) method was recently developed, where individual similarity networks from individual types of omics data were iteratively updated using information from other types of omics data through a nonlinear combination method (B. Wang, Mezlini, et al., 2014). We recently developed an annotation boosted SNF to further improve the clustering performance when association signals were used as weights on different types of omics data before fusing them into a fused similarity matrix (Ruan et al., 2019). Here we applied this nonlinear combination method to integrate patient similarity matrices from different domains of patient data.

Specifically, for the mth domain of patient data, we first define a global similarity network $P^{(m)}$ and a local similarity network $Q^{(m)}$ using the patient similarity network $S^{(m)}$ defined in Step 1. The entries of the global similarity network $P^{(m)}$ are defined as the normalized entries in $S^{(m)}$ introduced in Equation 2.3, and the entries of the local similarity network $Q^{(m)}$ are defined as the normalized similarities between patient i and his/her neighbors N_i as defined in Step 1, and 0 between patient i and subjects outside of his/her neighbors N_i. This local similarity network $Q^{(m)}$ is constructed with an assumption that local similarities might be more reliable than remote ones.

The global similarity networks $P^{(m)}, m = 1, ..., M$ for M domains of patient data are then smoothened through the parallel interchanging diffusion process (B. Wang, Mezlini, et al., 2014) that updates the global similarity network $P^{(m)}$ using the local similarity networks $Q^{(m)}$ and the global similarity networks of other domains of patient data. Consider the case where there are only two domains of patient data. We have global similarity networks $P^{(1)}$, $P^{(2)}$ and local similarity networks $Q^{(1)}$, $Q^{(2)}$, respectively. To update $P^{(1)}$, $P^{(2)}$ iteratively, let initial condition $P^{(1)}(t = 0) = P^{(1)}$ and $P^{(2)}(t = 0) = P^{(2)}$ for the first iteration, the diffusion process is described as follows:

$$P^{(1)}(t+1) = Q^{(1)} \times P^{(2)}(t) \times \left(Q^{(1)}\right)^T \tag{2.4}$$

$$P^{(2)}(t+1) = Q^{(2)} \times P^{(1)}(t) \times \left(Q^{(2)}\right)^T \tag{2.5}$$

After t iterations, the integrated similarity network is calculated as the average of the two updated globe similarity networks $P^{(fused)} = \left(P^{(1)}(t) + P^{(2)}(t)\right)/2$. When there are more than

two domains of patient data, the diffusion process Equations 2.4 and 2.5 can be expressed as:

$$P^{(m)} = Q^{(m)} \times \frac{\sum_{k \neq m} P^{(k)}}{M-1} \times \left(Q^{(m)}\right)^T, m = 1, ..., M \tag{2.6}$$

Step 3: Building a prediction tool

With a training set where samples' binary outcomes of interest are known (e.g., case vs. control), our goal is to predict the binary outcomes for samples in a test set. To do so, we first calculate the fused patient similarity network $P^{(fused)}$ with all samples in the training and test sets together. Note that in calculating $P^{(fused)}$, neighbors for the local similarity network of a test sample are from the training data only. Hence similarity measures of a test sample would not be affected by other test samples.

Using the training set, we assign a similarity t-score to each sample in the training set using the leave-one-out method as follows. For a case sample in a training set with n_1 cases and n_2 controls, this case sample's similarity t-score is the two-sample t-test statistic comparing its similarity with other $n_1 - 1$ cases, and its similarity with all n_2 controls. Similarly, for a control sample in a training set with n_1 cases and n_2 controls, this control sample's similarity t-score is the two-sample t-test statistic comparing its similarity with all n_1 cases, and its similarity with other $n_2 - 1$ controls. After all samples in the training set are assigned a similarity t-score, we fit a simple logistic regression of the known case-control status on the assigned similarity t-scores. This logistic regression model serves as a similarity-based prediction model, i.e., a classifier that can be used to predict test samples' case-control status.

To predict case-control status of samples in the test set using the similarity-based prediction model, we similarly assign samples in the test set a similarity t-score from a two-sample t-test statistic comparing similarities between a test sample and all n_1 cases in the training set, and similarities between the test sample and all n_2 controls in the training set. After assigning similarity t-scores to the test samples, we can then calculate the probability of each test sample being a case using the fitted logistic regression classifier.

We evaluate our method using receiver operating characteristic (ROC) curve and area under the curve (AUC), F_1-score, F_2-score, as well as the recall and precision. The F_β score is a weighted harmonic mean of recall and precision with the formula:

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{2.7}$$

where β represent relative importance such that recall is considered β times as important as precision. F_1-score considers equal weights for recall and precision, while F_2-score considers recall twice as important as precision (Grobelnik, 1999). The threshold for the probability of being a case is set at 0.5 for F_1-score, F_2-score, recall and precision.

2.2.2 Simulation studies

We conducted extensive simulation studies to investigate the prediction performance of PsDF and compared to that of the three competing methods.

Simulation settings

In our simulation studies, we considered three different simulated data domains. Domains 1 and 2 have a number of binary features and mimic typical domains based on medical records, e.g., indicating if a drug exposure or a medical procedure is recorded in EHR. Domain 3 has a single continuous feature. We considered a binary outcome. For each simulated binary feature, we generated measures from two different Binomial distributions for cases and controls with probability of success p_{case} and $p_{control}$, respectively. All features in Domain 1 are set to have large signals with the same p_{case1} and $p_{control1}$. All features in Domain 2 are set to have small signals with the same p_{case2} and $p_{control2}$. We considered two simulation scenarios.

The first scenario investigates the impact of imbalance among different domains where we fixed the number of features in Domain 1 at 5, and ranged number of features in Domain 2 from 10 to 200. We set $p_{control1} = 0.1$, and $p_{case1} = 0.4$ for features of large effect sizes in Domain 1 and

$p_{control2} = 0.1$ and $p_{case2} = 0.12$ for features of small effect sizes in Domain 2. For the continuous feature in Domain 3, we generated measures from a Gaussian distribution with means 0.1 and 0 for cases and controls, and with the standard deviation (SD) 1 for both groups.

The second scenario investigates the influence of nonlinear signals, such as variance signals, where we ranged SD of the single feature in Domain 3 for cases from 0.2 to 1 when the effect sizes of all other features are the same as those in the first scenario. In addition, the number of features in Domain 2 is fixed at 10. In all simulation settings, we set the scaling parameter μ as 0.5, and the size of neighbors K as $N/2$ in Equation 2.2, where N is the sample size of a training set.

The above simulation settings with $p_{case} > p_{control}$ mimic data in real EHR domains when cases usually have more records than controls. In order to evaluate PsDF more comprehensively, we also considered parallel scenarios when $p_{case} < p_{control}$, i.e., when binary features are more frequent in controls than in cases, and we set $p_{control1} = 0.5$, $p_{case1} = 0.2$ for features in Domain 1 and $p_{control2} = 0.5$, $p_{case2} = 0.48$ for features in Domain 2.

We simulated a population pool of 5,000 cases and 5,000 controls. We considered two designs with 1:1 case/control ratio and 1:5 case/control ratio when the latter with more controls is more common in EHR data. With the 1:1 case/control ratio, we randomly selected 200 cases and 200 controls as the training set, and another 100 cases and 100 controls as the test set. Therefore, for the training set, Domain 1 is a data matrix $Y^{400\times p_1}$, Domain 2 is a data matrix $Y^{400\times p_2}$ and Domain 3 is a data matrix $Y^{400\times 1}$. Data matrices of test set are similar. With the 1:5 case/control ratio, we randomly selected 200 cases and 1,000 controls as the training set, and another 100 cases and 500 controls as the test set. We repeated the simulation procedure 100 times and obtained average AUCs, F_1-scores, F_2-scores, recalls and precisions with their 95% confidence intervals (CIs).

2.2.3 Comparison methods

We considered three comparison methods, a random forest classifier and a logistic regression both using all features in Domains 1, 2 and 3 as predictors to classify case and control groups, and a naïve similarity method where the patient similarity matrix $S^{N\times N}$ is calculated using concatenated

features in Domains 1, 2, and 3 with applying Equations 2.1 and 2.2 and the same prediction step as described in PsDF Step 3.

2.3 Results

2.3.1 Simulation studies

We show the average AUCs, F_1-scores, F_2-scores, recalls and precisions when the threshold for the probability of being a case is 0.5 and their corresponding 95% CIs on test sets for the two simulation scenarios, 1) increasing the number of features in Domain 2, and 2) increasing the effect size of the variance signal of the single feature in Domain 3. These two simulation scenarios were done in parallel for two different settings, when cases have more EHR records than controls and when cases have fewer EHR records than controls. Finally, all simulation studies were done for the 1:1 case/control ratio (Figure 2.2) and the 1:5 case/control ratio (Figure 2.3).

For the 1:1 case/control ratio and when cases have more EHR records than controls (Figure 2.2A), when the number of signal features in Domain 2 is comparable (the number of signal features in Domain 2 is 10) to that in Domain 1, all four methods have similar prediction performance in terms of AUCs and F_1-scores, with PsDF having slightly higher F_1-scores. In addition, PsDF has the highest F_2-scores because its recalls are also the highest among four methods and F_2-score weights more on the recall. As the number of signal features in Domain 2 increases from 10 to 200, AUCs, F_1-scores, and F_2-scores of the logistic regression method quickly decrease as expected, while those of the other three methods are hardly affected. This is because regression methods often require certain ratios of sample size to a number of features in models in order to achieve a good model fit. When cases have fewer EHR records than controls (Figure 2.2B), AUCs, F_1-scores, and F_2-scores of the naïve similarity method slowly decrease. This is because the effective effect sizes of signal features in Domain 2 ($p_{control2} = 0.5$ and $p_{case2} = 0.48$) are much smaller than that in Figure 2.2A ($p_{control2} = 0.1$ and $p_{case2} = 0.12$), they become "noise features" to some extent. Therefore, for the naïve similarity method, as Domains 1 and 2 become more imbalanced, the contribution of 5 signal features with strong effect sizes in Domain 1 become weakened with

the increasing number of very small effect size features in Domain 2. Thus, the performance of the naïve similarity method becomes worse as the total number of features increases across all three domains combined. Note that, PsDF and random forest are not affected.

When we increase the effect size of the single continuous feature with variance signal in Domain 3 while keeping the number of features in Domains 1 and 2 at 5 and 10, respectively (Figure 2.2A and Figure 2.2B), AUCs, F_1-scores and F_2-scores of PsDF and random forest increase rapidly, while those of the logistic regression and the naïve similarity methods do not change much. This is also expected as regression methods cannot capture variance signals and the single variance signal feature in Domain 3 will be similarly diluted in the concatenated pool of signal features across the three domains for the naïve similarity method.

For the 1:5 case/control ratio (Figure 2.3), the overall patterns are similar to that of the 1:1 case/control ratio, with two noticeable differences: 1) When increasing the number of signal features in Domain 2, the performance of logistic regression does not decrease too much. This is because the total number of training samples (200 cases and 1,000 controls) are large enough. 2) F_1-scores and F_2-scores of random forest decrease quickly when the number of signal features in Domain 2 increases as random forest tends to classify almost all samples as controls. When cases have fewer EHR records than controls (Figure 2.3B), i.e., when the effective effect sizes of signal features are even smaller, or very close to the noises, random forest may classify very few samples or even 0 samples as cases, resulting in a very low recall or even 0 recall, thus and an unavailable precision. The performance of PsDF is not affected by the 1:5 case/control ratio.

2.3.2 Clinical study - ESKD prediction tools

More than 47,000 Americans die from chronic kidney disease (CKD) annually (J. Xu et al., 2016), yet the disease often has no symptoms in early stages and frequently goes undetected until it is advanced. In fact, less than 10% of patients affected with early CKD (stages 1-3), and only half (52%) of those with severe CKD (stage 4) are aware of having a kidney problem (Dharmarajan et al., 2017). Kidney disease usually gets worse over time, and although treatment has been shown to

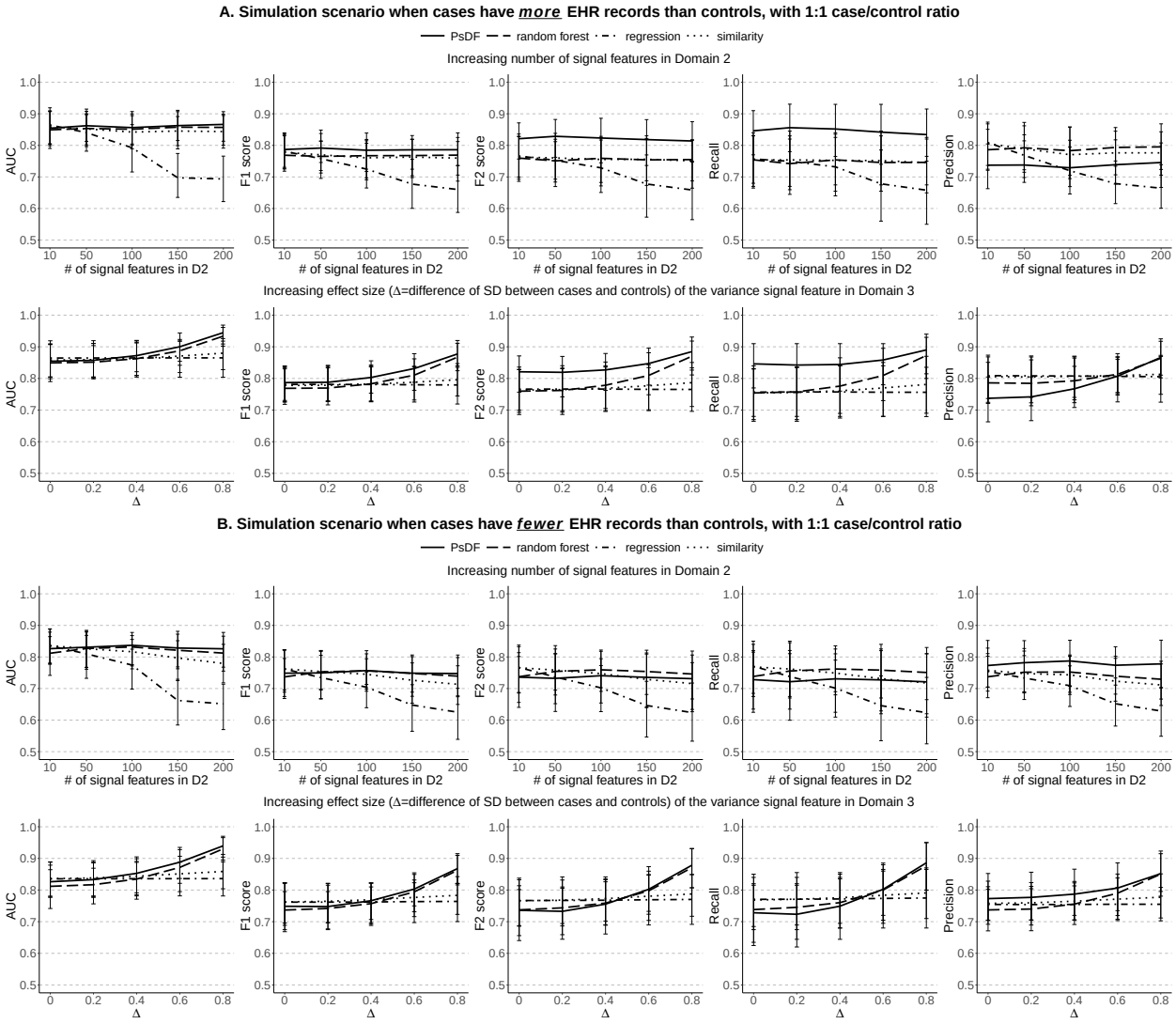

Figure 2.2: With the 1:1 case/control ratio, simulation results of prediction performance of the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method, under two simulation scenarios: 1) increasing number of signal features in Domain 2, and 2) increasing effect size of the variance signal feature in Domain 3. Part A displays results when cases have more EHR records than controls. Part B displays results when cases have fewer EHR records than controls.

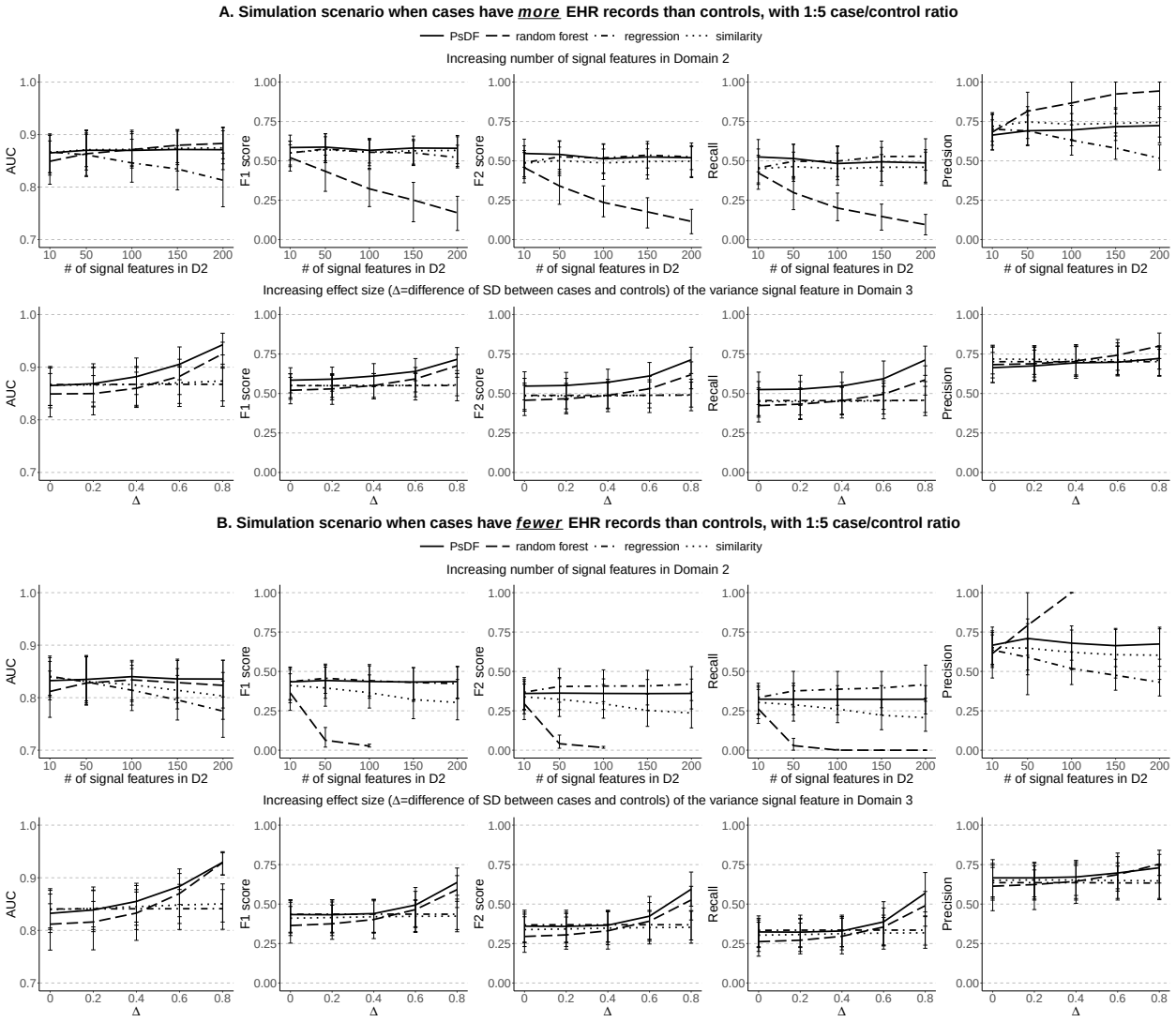

Figure 2.3: With the 1:5 case/control ratio, simulation results of prediction performance of the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method, under two simulation scenarios: 1) increasing number of signal features in Domain 2, and 2) increasing effect size of the variance signal feature in Domain 3. Part A displays results when cases have more EHR records than controls. Part B displays results when cases have fewer EHR records than controls.

delay progression, any preventive strategies are only effective when implemented early. When the kidney dysfunction reaches the level of "failure" (i.e., end stage kidney disease, ESKD), dialysis or kidney transplant are needed for survival. This state is irreversible and associated with accelerated cardiovascular disease and high mortality (Go et al., 2004). This highlights a great need for early diagnosis of CKD and identification of patients at risk of progression to ESKD, motivating our use of CKD as the first case study for PsDF.

We applied the PsDF algorithm and the three competing methods to build ESKD prediction tools and compared their performance. We predicted incident ESKD between 2006-2016 using comprehensive EHR data collected in years 2006 and prior. We used two different inclusion criteria to define eligible patients, a less stringent criterion that only requires patients to have demographic domain; and a more stringent criterion that requires patients to have demographic domain as well as records across all four EHR domains.

We conducted a sensitivity analysis to evaluate the robustness of PsDF and the three competing methods by randomly masking a percentage of observed EHR records in the test set by setting them to "missing". We masked 5-50% records in the test set with the increment of 5% to generate new test sets with more missing data than that in the training set.

EHR data preprocessing for ESKD prediction

We defined ESKD as chronic kidney disease (CKD) stage 5 (estimated glomerular filtration rate < 15 mL/min/1.73m^2) or CKD requiring kidney transplant, or any form of chronic dialysis. Among all patients in the CUIMC EHR data warehouse as of year 2006, 386,297 patients had sufficient data to define their CKD status. Among those, there were a total of 11,802 cases of ESKD and 374,495 non-ESKD patients (normal renal function or CKD stage 1-4). Among 374,495 non-ESKD patients, as of year 2016, 2,080 developed incident ESKD between 2006 and 2016, 353,295 remained non-ESKD, and the remaining 19,120 had status unknown. We considered those 2,080 patients who were non-ESKD in 2006 but reached ESKD before 2016 as our incident ESKD cases, and those 353,295 non-ESKD patients who remained non-ESKD between 2006 and 2016 as our

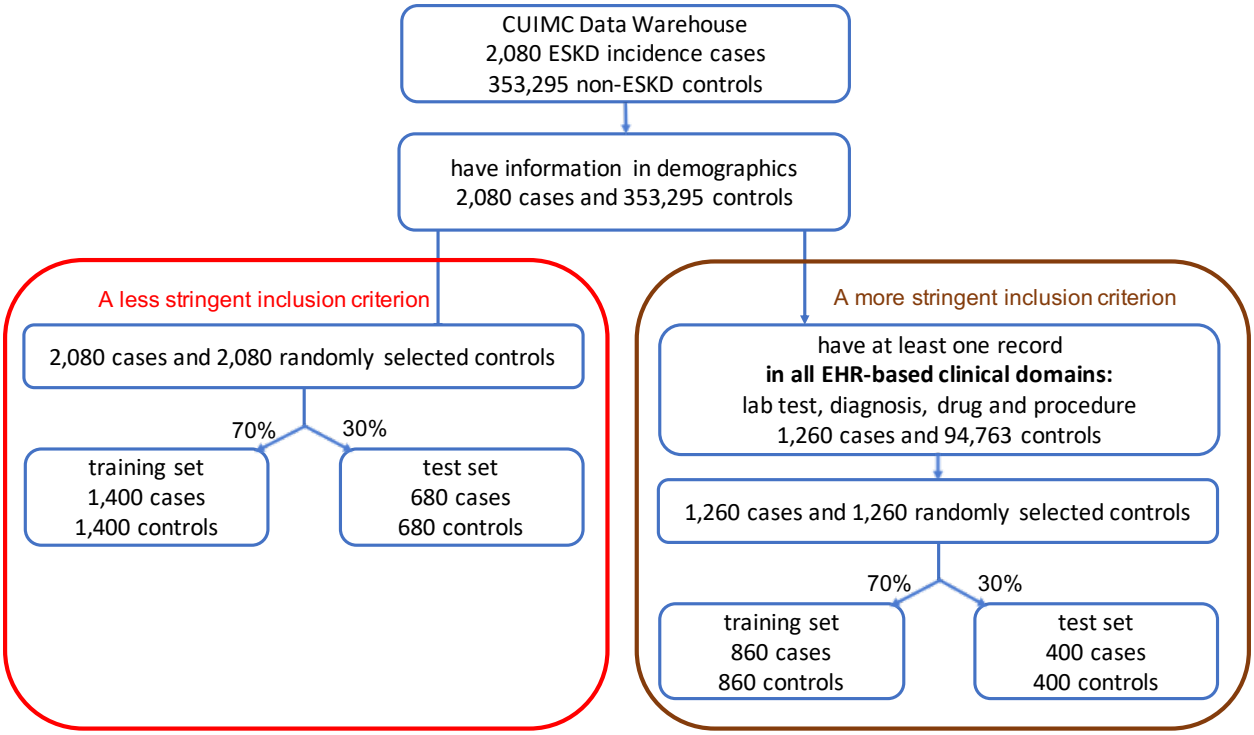

Figure 2.4: ESKD data preprocessing pipeline with two different inclusion criteria to define eligible patients.

controls. Our data processing pipeline is summarized in Figure 2.4. The comprehensive patient data included: 1) laboratory tests, 2) ICD based diagnosis history, 3) drug exposures, 4) medical procedures, and 5) demographic information with gender and race (white vs. non-white).

After requiring all patients to have demographic data, we had 2,080 ESKD cases and 353,295 non-ESKD controls. We then applied two different inclusion criteria on the four EHR domains to define eligible patients in the study: 1) the less stringent inclusion criterion which does not have any requirement on EHR domains; 2) the more stringent inclusion criterion which requires patients to have records across all four EHR domains. Figure 2.4 displays the data preprocessing pipeline and the final sample sizes with the two inclusion criteria.

A less stringent inclusion criterion

Patients were included if they had demographic information, resulting in 2,080 ESKD patients and 353,295 non-ESKD controls. We then randomly selected 2,080 patients among 353,295 non-ESKD controls to create a balanced case control design, as it is known that a balanced design

helps to reduce variances of estimated parameters in logistic regression models (King & Zeng, 2001). We split 2,080 ESKD cases and 2,080 non-ESKD controls into two cohorts, one as the training set with 1,400 ESKD cases and 1,400 non-ESKD controls, the other as the test set with 680 ESKD cases and 680 non-ESKD controls to test the prediction performance of PsDF and the three competing methods.

A more stringent inclusion criterion

Patients were included if they had demographic information as well as records in all four EHR domains, resulting in 1,260 ESKD patients and 94,763 non-ESKD controls. We then randomly selected 1,260 patients among the 94,763 non-ESKD controls to make a balanced case control design. We similarly split 1,260 ESKD cases and 1,260 non-ESKD controls into two cohorts, one as the training set with 860 ESKD cases and 860 non-ESKD controls, the other as the test set with 400 ESKD cases and 400 non-ESKD controls.

In order to investigate the model performance under an unbalanced case-control design, we also considered a 1:5 case/control ratio. That is, in addition to the previously selected controls, we randomly selected another 8,320 controls (four times of 2,080 cases) for the less stringent inclusion criterion, and another 5,040 controls (four times of 1,260 cases) for the more stringent inclusion criterion. For each criterion, we split these additional control samples into two groups with ratio 70% and 30%, then added them into the training set and test set accordingly.

Feature selection using LASSO regression and random forest

Because of the large number of features in every EHR domain, we included a screening step to pre-select potentially relevant features using LASSO regression and random forest in order to capture both linear and nonlinear features for prediction. We used the training set with 1:1 case/control ratio for this step.

We applied the stability selection using LASSO regression on each domain separately. Specifically, we resampled a subset of size $N/2$ of the training set with sample size N without replacement. We then applied LASSO regression on the subset and obtained a set of selected features

of non-zero regression coefficients. We repeated this subsampling 1,000 times and obtained the selection probability for each feature out of the 1,000 subsampling. We then selected features with selection probability greater than 0.6. With the training set defined by the less stringent inclusion criterion, we selected 19 features out of 1,123 laboratory tests, 23 of 7,980 diagnostic history features, 18 of 3,936 drug exposure features, 34 of 6,324 medical procedure features, as well as gender and race out of demographic variables. With the training set defined by the more stringent inclusion criterion, we selected 26 laboratory tests, 26 diagnostic history, 26 drug exposures, 23 medical procedures, as well as gender and race.

We then applied random forest on the training set to select features with nonlinear signals for each domain separately and selected features with high importance, defined as mean decrease accuracy. We used a threshold of greater than 0.1 for the importance measure. With the training set defined by the less stringent inclusion criterion, we selected 23 laboratory tests, 66 diagnostic history, 45 drug exposures, 42 medical procedures, and gender. With the training set defined by the more stringent inclusion criterion, we selected 24 laboratory tests, 31 diagnostic history, 31 drug exposures, 21 medical procedures, as well as gender and race.

We unionized the features selected by LASSO and random forest, which led to 204 features in total for the dataset of less stringent inclusion criterion and 145 features in total for the dataset of more stringent inclusion criterion.

Comparison of the four prediction methods

To compare the prediction performance of the four methods using the test set, we applied bootstrapping 1,000 times on the test set and obtained average AUCs, F_1-scores, F_2-scores, recalls and precisions when the threshold for the probability of being a case is set at 0.5, as well as their 95% CIs. In addition, we conducted a sensitivity analysis where we masked certain percentages of observations to investigate the robustness of PsDF and the three competing methods to missing data, as previously explained. Figure 2.5 summarizes prediction performance for the ESKD prediction tools from two inclusion criteria for the 1:5 case/control ratio. The results for the 1:1 case/control

ratio are included in the Appendix Section A.1. The results are very similar to that of the 1:5 case/control ratio.

In general, both PsDF and random forest outperform logistic regression and the naïve similarity method in terms of AUCs (Figure 2.5). Without missingness ($p_{mask} = 0$), AUCs of PsDF, random forest and logistic regression are comparably high, at approximately 0.85 with overlapping 95% CIs. When the robustness of the three methods is tested against the variable degree of missingness, AUCs drop dramatically for the logistic regression and the naïve similarity methods, while F_1-scores and F_2-scores drop quickly for random forest, with increasing masking percentage (p_{mask}). In contrast, AUCs, F_1-scores and F_2-scores are all relatively stable for the PsDF method, demonstrating a clear advantage of this method over the other three competing methods.

We also note that the ESKD prediction tool developed by PsDF has higher recalls and lower precisions than those of the other three competing methods when the threshold for the probability of being a case is set at 0.5. Because ESKD cases usually have more EHR records than non-ESKD controls, this pattern resembles the one observed in the simulation studies when cases were set to have more EHR records than controls (Figure 2.2A). We also observed decrease in recalls with increasing missingness for all four methods, however, the recalls of PsDF decrease much slower than those of the other three methods, while the recalls of random forest decrease dramatically, similar as the patterns in simulation studies.

As there are limited geocoding information available for some of the EHR patients, for demonstration purposes that PsDF can fuse all available domains, we repeated the construction of the ESKD prediction tools including the geocoding domain. We updated the samples selection for the training and test sets accordingly. There are two continuous variables available for the geocoding domain, median household income in dollars and distance to the nearest major road in meters. Other five domains are the same as described above. The patterns of AUCs, F_1-scores, F_2-scores, recalls and precisions are similar to those with 5 domains. Full description of the procedure and results is included in the Appendix Section A.2.

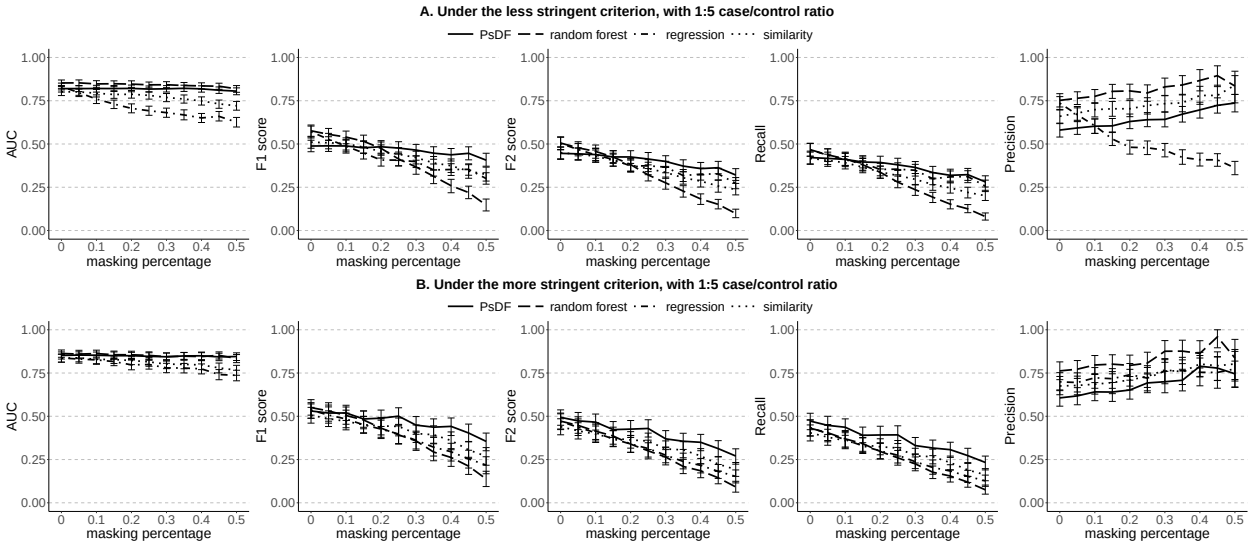

Figure 2.5: With the 1:5 case/control ratio, prediction performance of the ESKD prediction tools built by the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method when the masking percentage p_{mask} increases, under two different inclusion criteria: A) the less stringent criterion, and B) the more stringent criterion.

2.3.3 Clinical study - AS prediction tools

Similar to kidney disease, the natural history of aortic stenosis (AS) progresses through a prolonged asymptomatic period prior to the development of symptomatic disease that requires valve replacement. Although there is an average rate of reduction of valve area quoted from epidemiologic studies, there are some patients who undergo rapid progression of disease and others who have minimal to no progression over a similar time frame. The targeted use of surveillance ultrasound to monitor progression of AS and to determine when valve replacement should occur could reduce unnecessary medical spending and help direct limited resources to patients who need them most. The application of PsDF to identify patients at high risk of disease progression may further facilitate planning of a valve replacement procedure. We therefore applied PsDF and the three competing methods to build AS prediction tools. The patterns of AUCs, F_1-scores, F_2-scores, recalls and precisions are similar to those observed for the prediction of ESKD. Full description of the methods and results is included in the Appendix Section A.3.

2.4 Discussion

We developed Patient similarity based on Domain Fusion (PsDF), a novel framework for clinical outcome prediction using comprehensive patient data. The PsDF method integrates similarity information from multiple data domains into a comprehensive similarity measurement that can be subsequently used to predict important clinical outcomes. In contrast to the similarity-based methods based on concatenated data, our fusion method allows for highly unbalanced data domains to be treated equally, and prevents any domain with a large number of features from dominating the prediction. Moreover, as a similarity-based method, PsDF naturally captures nonlinear signals, such as variance-based signals, and does not require a certain ratio of sample size to the number of features that is required for regression-based models. We demonstrate that PsDF is highly flexible, scalable, and makes use of the entirety of patient's data (EHR-based as well as non-EHR-based) to define comprehensive similarity. With extensive simulation studies, we demonstrate an improved prediction performance of PsDF over the competing methods, including random forest, logistic regression and naïve similarity methods. In the presence of nonlinear signals and when domains with unbalanced sizes exist, PsDF outperforms the competing methods through its ability to preserve strong signals, accumulate weak signals, and capture nonlinear effects.

In two clinical application studies, we also demonstrate that PsDF is more robust to random missingness compared to random forest, logistic regression or naïve similarity methods. This is an important advantage, given that missing data is a ubiquitous property of the real life EHR data. This advantage stems from the fact that PsDF integrates similarity information across different domains and performs prediction based on integrated relative similarity between a sample in the test set and all samples in the training set. Even though the masking procedure may change distributions of features in the test set, the relative similarity to the training set may not change much. On the other hand, random forest tends to classify almost all test samples as controls, especially with an unbalanced case-control design with more controls than cases. Logistic regression-based methods rely heavily on parameter estimates for selected features using the training set. When the features'

distributions in the test set are different from that in the training set, it is expected that the prediction performance of logistic regression would rapidly decrease. The naïve similarity method is also not expected to be robust to missingness, because the dilution of signal features with concatenation becomes even more severe when some observations are masked.

We want to emphasize that the features used in the two clinical studies were pre-selected by LASSO and random forest, which favor the two competing methods, i.e., logistic regressions and random forest. Close investigation of the selected features for the ESKD prediction tool and the AS prediction tool suggests that they are clinically reasonable. For example, in the ESKD prediction tools, "disorder of kidney and/or ureter", "biopsy of kidney" and "acute renal failure syndrome" were selected under both less/more stringent inclusion criterion. In the AS prediction tools, "aortic valve disorder", "cardiac complication" and "diagnostic ultrasound of heart" were selected under both less/more stringent inclusion criterion.

One limitation of the current study is that we coded all features in EHR-based domains to be binary, indicating the presence or absence of a record. We did not use cumulative counts or continuous measures of certain features, which likely led to some information loss. Another limitation is that we did not use longitudinal information embedded in patient records, nor did we consider different visit types (e.g., hospital versus ambulatory). We are currently working on extending the PsDF framework in order to make full usage of such information. We want to emphasize that the prediction performance could be further enhanced if data from more patient domains becomes available in the future, such as genetic or exposome data. The success of our two clinical application studies suggests that the framework of PsDF is highly flexible, scalable, and generalizable, and thus this method has a great potential in developing new patient similarity-based clinical prediction tools.

Chapter 3: PheW^{2}P2V - a phenome-wide prediction framework with weighted patient representations using electronic health records

3.1 Introduction

The increasing adoption of electronic health records (EHRs) brings opportunities to develop new computational predictive tools (Agrawal & Prabakaran, 2020; Dash et al., 2019; Shivade et al., 2014). Conventional machine learning approaches such as regression-based, bagging, or boosting methods have been widely used to predict clinical outcomes such as heart failure, type 2 diabetes mellitus, hypertension, and others (Anderson et al., 2016; Henderson et al., 2018; Steele et al., 2018; J. Wu et al., 2010; Zein et al., 2021). We recently developed a flexible similarity-based algorithm and applied it to predict end stage kidney disease and severe aortic stenosis (J. Guo et al., 2021). With these conventional methods, usually one prediction tool is trained for one outcome, i.e., they are outcome-specific, and only labeled data are used to train the model, i.e., they are fully supervised. In addition, these conventional methods usually take data that are well-structured without missing values.

Deep learning algorithms for natural language processing (NLP) have also been used for clinical decision making with EHR, because sequences of medical records of patients in EHR are similar to sequences of words in text documents. Historically developed for NLP tasks such as machine translation to fully use sequence information, recurrent neural network (RNN) (Hochreiter & Schmidhuber, 1997; Rumelhart et al., 1986) has been widely used in EHR to predict health outcomes. With the word2vec algorithm being introduced (Mikolov et al., 2013) in 2013, which represents words with numeric vectors, medical records embeddings can be pre-trained and combined with prediction models such as RNN or logistic models, either in a two-step fashion, i.e., embedding plus prediction, or in a single-step fashion. In a single-step approach, RNN models use

pre-trained embeddings to fine tune them together with RNN parameters to predict outcomes such as diagnoses, or readmissions (Ashfaq et al., 2019; Che et al., 2018; Choi et al., 2016; J. Zhang et al., 2018). Thus, one-step models are outcome-specific, fully-supervised, and are thus computationally intensive. On the other hand, two-step models, i.e., embedding plus prediction, are not outcome-specific, as embedding is done once and is combined with a prediction model to predict outcomes. They are not fully supervised as unlabeled data can be used for embedding. Two-step models are thus computationally efficient. Farhan et al. proposed a two-step model (Farhan et al., 2016), where a patient's sequence of medical records was represented by summing up numeric vectors of his(her) medical records and was subsequently used to predict the patient's risks of many diagnoses. However, the prediction performance is only slightly better than that of logistic regressions because all medical records were treated equally regardless of the outcome of interest.

To avoid RNN and improve computation, the Transformer model (Vaswani et al., 2017) was developed, which uses a position embedding and self-attention layers being parameterized by three additional weight matrices to capture relative contribution of other words in a sentence. Transformer conducts predictions in a one-step fashion where embeddings are fine-tuned for a specific outcome. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) was subsequently developed to improve Transformer through pre-training a Transformer encoder by predicting randomly masked words. BERT can be combined with different prediction models, such as RNNs or Transformer decoders, either in a two-step fashion, or a one-step fashion. BERT has been applied in EHR as well. The Med-BERT model (Rasmy et al., 2021) was pre-trained on a large external dataset with 28 million patients, and the model was then fine-tuned using RNN to predict two diseases, heart failure among diabetic patients and onset of pancreatic cancer. However, the prediction gains of Med-BERT through pre-training using such a large external dataset are minimal compared to the computational cost.

In this work, we aim to achieve the goal of phenome-wide predictions while maintaining computational efficiency and good prediction performance for individual phenotypes. The prospective phenome-wide predictions could be useful as a screening tool to identify and flag patients with

high-risk conditions in early stages which may be missed overwise. We propose PheW^{2}P2V, a general Phenome-wide prediction framework that uses Weighted Patient Vectors. PheW^{2}P2V is the first phenome-wide prediction framework that alleviates the limitations of being outcome-specific and thus computationally intensive by taking a two-step procedure, i.e., embedding plus prediction. To maintain good prediction performance, PheW^{2}P2V uses a novel weighting scheme on patient embeddings so that predictions based on patient embeddings are tailored to individual phenotypes. Since diagnosis concepts in EHR are usually coded using International Classification of Disease (ICD) terminology, which is designed for billing and administrative functions but not for case-control studies (Denny et al., 2010), PheW^{2}P2V first maps patients' ICD codes to clinical disease phenotypes called phenotype codes (phecodes). Phecodes are originally developed for phenome-wide association studies (PheWAS), where patients' phenotypes are identified by grouping ICD codes that represent common etiologies, with a purpose of reducing the redundancy in ICD codes and better defining cases and controls (Denny et al., 2013; Denny et al., 2010). To predict a clinical disease phenotype in the phenome, after generating medical concepts embeddings using word2vec, PheW^{2}P2V represents each patient as a single numeric patient vector where his(her) medical concepts that are more correlated with the phenotype of interest will be upweighted. The patient vector is then used to predict the incidence risk of the phenotype. Unlike the one-step model where embeddings are fine-tuned for one outcome of interest, PheW^{2}P2V introduces weights on medical concepts to improve computational efficiency while tailoring predictions to the specific phenotype of interest to maintain good phenome-wide prediction performance. Unlabeled data can also be used for medical concepts embeddings, which makes PheW^{2}P2V not fully supervised.

Using simulation studies, we showed an improved prediction of PheW^{2}P2V over that of four baseline methods including a regression-based model, a random forest classifier, a gradient boosted tree classifier, and the P2V model without weights (Farhan et al., 2016). We applied PheW^{2}P2V and baseline methods to the MIMIC-III database (Johnson et al., 2016) to predict patients' incidence risks of 942 phenotypes at the latest visit using medical records from past visits. We observed better predictions of PheW^{2}P2V consistently across most phenotypes over that of base-

line methods. We also demonstrated several clinical examples in which PheW^{2}P2V can predict less common conditions that could be diagnostically challenging or missed on a routine clinical work up, such as adrenal insufficiency or chronic pericarditis. Automated suggestions provided by PheW^{2}P2V that such conditions should be considered in the differential diagnosis and in the work up of high risk patients could be clinically impactful.

3.2 Methods and materials

3.2.1 The PheW^{2}P2V algorithm

The PheW^{2}P2V framework is illustrated in Figure 3.1 with four steps: (1) identifying case-control status of phenome-wide clinical disease phenotypes by mapping diagnosis ICD codes to phenotype codes (phecodes) and constructing patient sequences; (2) generating medical concepts embeddings using word2vec; (3) calculating weighted patient vectors with weights capturing correlations between past medical records and a phenotype of interest; and (4) conducting tailored phenome-wide predictions using weighted patient vectors to predict incidence risks of individual patients.

Step 1: Phenome-wide case-control identification

Phenotype mapping using Phecode map (Denny et al., 2010) (with R package "PheWAS" (Carroll et al., 2014)) aims to reduce the redundancy of ICD codes and more accurately define case-control status of clinical phenotypes, for the purpose of phenome-wide analysis. With phecode mapping, over 14,000 codes in the ICD-9 system were grouped into approximately 1,600 phenotype codes (phecodes) (Denny et al., 2013). Using the MIMIC-III database, for each patient's visit, PheW^{2}P2V maps patient's ICD-9 codes to phecode-defined "case groups", i.e., case status of meaningful clinical phenotypes. A list of disease-specific exclusion phecodes was also specified for each "case group". A patient without any ICD-9 code in this list is defined as the "control group". This mapping ensures that patients with comparable diseases are not categorized as controls. For example, a patient with an unknown arrhythmia cannot be considered as a control for

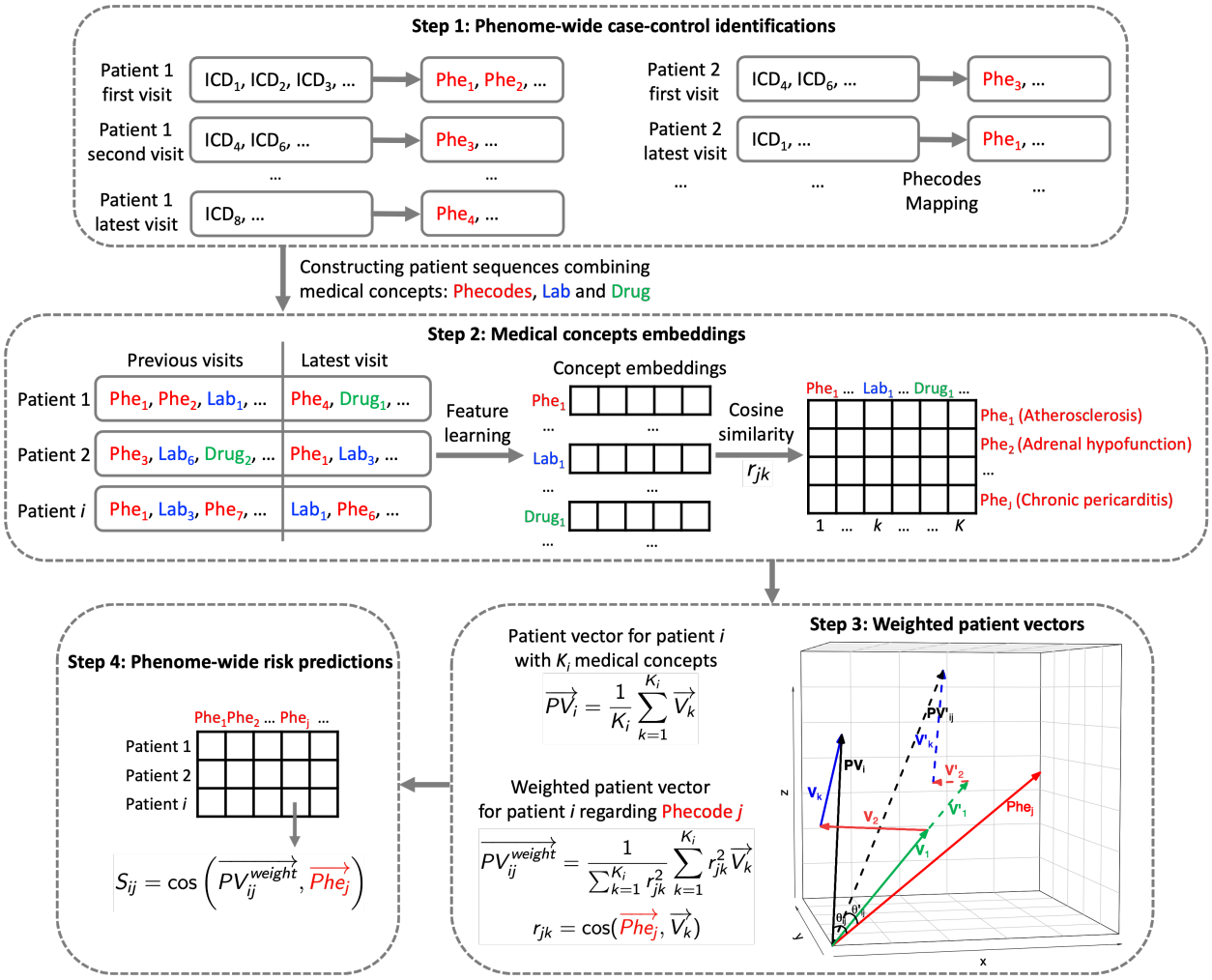

Figure 3.1: The workflow of the proposed PheW^{2}P2V framework.

atrial fibrillation (Carroll et al., 2014).

Step 2: Medical concepts embeddings

Word2vec is a widely used embedding algorithm in NLP that is computationally efficient (Mikolov et al., 2013). With a large corpus of text, it uses a neural network model to generate numeric vectors for unique words in the corpus. These numeric vectors have the same dimension and are thus embedded in a vector space. A patient's sequence of medical concepts that contains phenotype codes, lab test codes, etc, can be considered as a sentence with words. PheW^{2}P2V applies word2vec to learn numeric vectors for unique medical concepts. The clinical similarity between two medical concepts can then be captured by the cosine similarity between the two corresponding numeric vectors.

Step 3: Weighted patient vectors tailored for a phenotype

For each phenotype of interest, we calculate cosine similarities between the numeric vector of the phenotype and numeric vectors of all other medical concepts. Suppose an EHR database has K unique medical concepts, among which there are J unique phenotype codes. We build a correlation matrix with dimension $J \times K$ to capture correlations between J phenotypes of interest and K medical concepts. To conduct tailored predictions, a patient's past medical records are summarized into a numeric patient vector, which is a weighted average of numeric vectors of the patient's past medical concepts using phenotype-specific weights to up-weight medical concepts that are most relevant to the phenotype:

$$\begin{aligned} \overrightarrow{PV_{ij}^{\text{weight}}} &= \frac{1}{\sum_{k=1}^{K_i} r_{jk}^2} \sum_{k=1}^{K_i} r_{jk}^2 \overrightarrow{V_k}, \\ r_{jk} &= \text{cosine}\left(\overrightarrow{Phe_j}, \overrightarrow{V_k}\right). \end{aligned} \tag{3.1}$$

Here we calculate the patient vector $\overrightarrow{PV_{ij}^{\text{weight}}}$ for patient i tailored for phenotype j, where K_i is the total number of concepts of patient i, $\overrightarrow{V_k}$ is the vector representation of medical concept k,

$\overrightarrow{Phe_j}$ is the vector representation of phenotype j, and r_{jk} is the cosine similarity between medical concept k $\overrightarrow{V_k}$ and phenotype j $\overrightarrow{Phe_j}$ measuring their correlation. The weight of medical concept k tailored for $\overrightarrow{Phe_j}$ for patient i is defined as $\frac{r_{jk}^2}{\sum_{k=1}^{K_i} r_{jk}^2}$, where we treat negative and positive correlations equally.

Step 4: Phenome-wide risk predictions

To predict risk of phenotype j for a test sample t, we compute the test sample's patient vector $\overrightarrow{PV_{tj}^{\text{weight}}}$ using Equation 3.1 and and the following risk score, which is the cosine similarity between the patient vector and the phenotype vector:

$$S_{tj} = \text{cosine}\left(\overrightarrow{PV_{tj}^{\text{weight}}}, \overrightarrow{Phe_j}\right) \tag{3.2}$$

Risk score S_{tj} ranges from -1 to 1, with a higher value indicating higher incidence risk of the phenotype.

3.2.2 Comparison methods and evaluation metrics

We considered four comparison methods, i) a LASSO regression model, ii) a random forest classifier, iii) a gradient boosted tree classifier, and iv) P2V without weights. For LASSO and random forest, we used the default settings and implementations in the Python library "scikit-learn", where regularization strength of LASSO is C=1.0 and the number of trees in random forest is n=100 with Gini impurity as the split criterion. For the gradient boosted tree, we used the default settings in the Python library "xgboost" with the number of rounds n=100. For LASSO regression, random forest, and gradient boosted tree, we constructed a data matrix with rows representing patients and columns representing counts of medical concepts from past admissions before the latest admission.

We evaluate model performance using area under the receiver operating characteristic curve (AUC-ROC), max F_1-score, and area under the precision-recall curve (AUC-PR). An ROC curve

is created by plotting the true positive rate (also called sensitivity or recall) and false positive rate (1-specificity) at various discrimination thresholds (e.g., thresholds of risk score S_{tj} for test sample t and phenotype j for the proposed PheW^{2}P2V) to illustrate the prediction ability of a binary classifier. In general, an AUC-ROC of 0.5 suggests that the classifier is uninformative and assigns labels randomly. PR curves are similar to ROC curves, but with precision and recall as the axes. A random classifier has an AUC-PR (also called average precision) equal to the percentage of positive samples, i.e., the percentage of cases p_{case} for a phenotype. F_1-score is the harmonic mean of precision and recall $F_1 = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. A dummy classifier that identifies all samples as cases would have a $F_1 = (2 \times p_{case})/(p_{case} + 1)$. A discrimination threshold is needed to calculate the F_1-score. Since different methods might have different optimal thresholds for different phenotypes, we compute the maximum F_1-score across all possible discrimination thresholds for each prediction method.

3.2.3 The MIMIC-III database

MIMIC-III data preprocessing

We conducted a phenome-wide prediction using the MIMIC-III database. We aim to predict the incidence risks of individual phenotypes across the phenome at the latest visit using past medical records. MIMIC-III is a freely accessible critical care database (Johnson et al., 2016). We used medical concepts from three clinical domains, diagnoses history (ICD-9 codes), prescriptions, and lab tests. There are 46,520 unique patients in MIMIC-III, among which 39,001 had only one admission and 7,519 had at least two admissions. There are in total 58,951 admissions with 6,984 unique ICD-9 codes, 3,267 unique prescriptions, and 710 unique lab tests. For prescriptions and lab tests, we used binary information of whether a patient ever had been prescribed a specific drug during an admission and whether a patient ever had a specific lab test during an admission. We did not use continuous variables. With phecode mapping, the 6,984 ICD-9 codes across all 58,951 admissions were mapped to 1,693 phenotype codes in Step 1 of the proposed PheW^{2}P2V. To have good numeric representations of medical concepts using word2vec, we removed extremely rare medical

Table 3.1: Summary of the MIMIC-III database after data processing.

		MIMIC-III database
Admissions		58,951
Unique patients		46,520
Patients with only one admission		39,001
Patients with at least two admissions		7,519
Unique medical concepts excluding rare ones	Phenotype codes	956
	Prescriptions	1,348
	Lab tests	490
Phenotype codes for prediction (prevalence $\geq 0.05\%$)		942

concepts who appeared in fewer than 50 admissions across all 58,951 admissions. Similar procedures were taken by other studies with medical concept representations using word2vec (Farhan et al., 2016). After these steps, we have 956 unique phenotype codes, 1,348 unique prescriptions, and 490 unique lab tests. We used these medical concepts to construct patient sequences for individual patients, which are time-sorted (sorted by admissions) sequences of medical concepts. Table 3.1 summarizes information of patients and medical concepts in the MIMIC-III data after processing.

Incident cases identification for phenome-wide predictions

Our goal is to predict patients' incidence risks of phenome-wide phenotypes at the latest visit using patients' medical history from past visits. We define incident cases at the latest visit as patients who (1) had at least two visits, (2) were identified as cases of a phenotype at the latest visit, and (3) had never been identified as the case of the phenotype in past visits. Valid controls are patients who met the conditions (1) and (3) and were identified as controls of the phenotype at the latest visit. For each phenotype, incident cases and valid controls are labeled subjects, while other patients are unlabeled subjects (including patients with one visit, and patients being neither incident cases nor valid controls) which can be used in medical concept embeddings. We calculated phenotype prevalence as the percent of incident cases among all labeled subjects at the latest visit. Among 956 phenotype codes, 14 have prevalence less than 0.05% and were removed from phenome-wide predictions. We predicted the rest 942 phenotypes (Table 3.1).

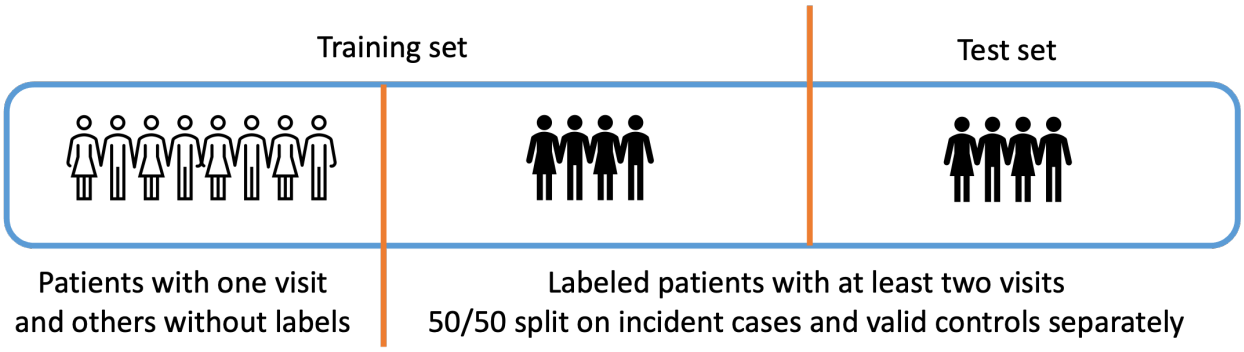

Figure 3.2: MIMIC-III sample splitting procedures for training and test samples.

With each phenotype for prediction, we half/half split labeled subjects into a training and a test set. Training samples included patients without labels and 50% of labeled patients, while test samples are the other 50% labeled patients (Figure 3.2). We repeated the half/half random training/test split 10 times to obtain average AUC-ROC, AUC-PR, and max F_1-score in test sets. Note that for different phenotypes, there are different numbers of labeled subjects.

3.3 Results

3.3.1 Simulation studies

We conducted simulation studies to evaluate the prediction performance of PheW^{2}P2V and that of comparison methods. We considered binary present/absent medical concepts, i.e., binary phenotypes and binary predictors. Specifically, we simulated a population pool of 20,000 patients each with a binary phenotype concept C_0 and 150 binary medical concepts, including 10 signal predictor concepts (denoted as $C_1, C_2, \ldots, C_{10}$) that predict C_0 and 140 noise concepts (denoted as $C_{11}, C_{12}, \ldots, C_{150}$). Those 150 binary concepts were generated to be correlated with each other, to mimic potential correlations between medical concepts. Detailed data generation steps were included in Appendix Figure B.1. The outcome phenotype concept C_0 was generated using a logistic model with the 10 signal predictor concepts, where we set the same β coefficients for them and considered different association strengths ranging β from 0.2 to 0.8 with a grid of 0.1. We set the intercept so that the probability of having outcome concept C_0 is around 0.5. Therefore, there

will be roughly 10,000 cases and 10,000 controls in the population pool of 20,000 patients. As we do not consider temporal information in a patient sequence, we randomly shuffled medical concepts of each patient. From the population pool of 20,000 patients, we bootstrapped 200 samples as a training set and another 200 samples as a test set. To mimic phenotype prevalence in the MIMIC-III database, we set the case-control ratio as 1:19 (10 cases and 190 controls) in both training sets and test sets, to have a phenotype prevalence of 5%. We repeated this procedure 1,000 times and obtained prediction results from 1,000 test sets for all methods. We also considered other case-control ratios including 1:1, 3:7, and 1:9 (with a prevalence of 50%, 30%, and 10%), and included results in Appendix Section B.3.

We summarized medians, 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR for PheW^{2}P2V and comparison methods across 1,000 simulated test sets in Figure 3.3. We can see that AUC-ROC, max F_1-score, and AUC-PR of all five methods increase as β increases as expected. PheW^{2}P2V outperforms all comparison methods, especially when signal is weak. Results for different case-control ratios were summarized in Appendix Figure B.3, where similar patterns were observed. We observed a bigger improvement of PheW^{2}P2V over LASSO regression, random forest, and gradient boosted tree for rare phenotypes because the imbalance between cases and controls affects the prediction performance of regression-based and tree-based methods more (C. Chen et al., 2004; Q. Wu et al., 2014) than that of P2V methods.

We also conducted simulation studies to demonstrate that medical concept embeddings using word2vec can recover the association strength between a signal medical concept (an explanatory predictor) and a phenotype (an outcome). Results are included in Appendix Section B.2.

3.3.2 Phenome-wide predictions using the MIMIC-III database

Phenome-wide prediction results using the MIMIC-III database was summarized in Table 3.2 with medians, 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR from 10 train/test splits across all 942 phenotypes binned with 300 phenotypes ranked by prevalence. Across the phenome, PheW^{2}P2V has a median AUC-ROC 0.73 (competing methods have values

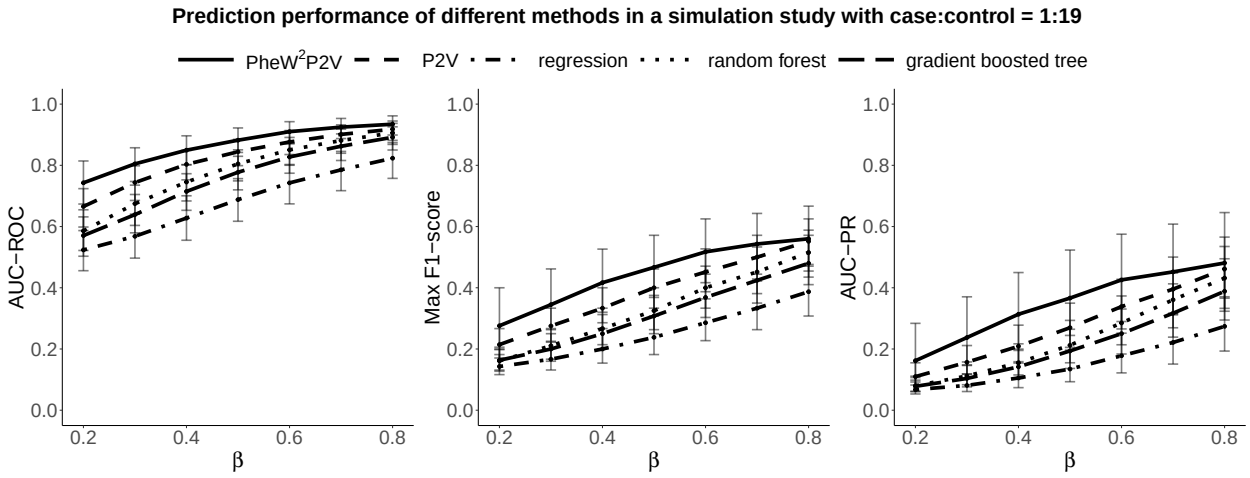

Figure 3.3: Simulation results of medians, 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR of the proposed PheW^{2}P2V, the LASSO regression, the random forest classifier, the gradient boosted tree classifier, and the unweighted version P2V with regression coefficient β ranging from 0.2 to 0.8, under the scenario of 1:19 case-control ratio.

$\leq$ 0.70), a median max F_1-score 0.20 (competing methods have values $\leq$ 0.19), and a median AUC-PR 0.10 (competing methods have values $\leq$ 0.10). There is a decreasing trend in prediction performance for all methods as phenotypes become rarer as expected. PheW^{2}P2V has bigger improvements over competing methods when phenotypes are rare, which is consistent with simulation results. Results in Table 3.2 were also plotted in Figure 3.4 for a better visualization where we can see that the proposed PheW^{2}P2V has the highest AUC-ROC, max F_1-score, and AUC-PR at almost all bins of phenotypes.

3.3.3 Examples of clinical disease phenotype predictions in the MIMIC-III database

We investigated individual phenotypes to understand the clinical significance of the proposed PheW^{2}P2V for phenome-wide predictions and highlighted 5 phenotypes from two different clinical categories with their prediction performance in Table 3.3. The first category includes common general medical conditions that are potentially preventable or treatable, such as atherosclerosis (phenotype code: 440) and diabetic retinopathy (phenotype code: 250.7). These disorders are frequently under-diagnosed and under-treated despite the fact that effective preventive and therapeutic strategies exist. PheW^{2}P2V has the best prediction performance for these conditions, for example,

Table 3.2: Medians and 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR of the 942 phenotypes binned by 300 from most to least prevalent phenotypes in the MIMIC-III database.

	Prevalence rank of phenotypes			
	1-300	**301-600**	**601-942**	**All**
Prevalence median (Q1, Q3)	0.042 (0.025, 0.075)	0.008 (0.006, 0.011)	0.003 (0.001, 0.003)	0.007 (0.003, 0.024)
AUC-ROC median (Q1, Q3)				
PheW^{2}P2V	0.78 (0.68, 0.87)	0.75 (0.68, 0.82)	0.69 (0.61, 0.80)	0.73 (0.65, 0.83)
P2V	0.71 (0.65, 0.79)	0.68 (0.63, 0.75)	0.65 (0.57, 0.74)	0.68 (0.62, 0.76)
Regression	0.72 (0.60, 0.85)	0.68 (0.59, 0.80)	0.59 (0.50, 0.78)	0.66 (0.57, 0.81)
Random forest	0.76 (0.65, 0.89)	0.70 (0.61, 0.81)	0.58 (0.53, 0.75)	0.69 (0.58, 0.82)
Gradient boosted tree	0.76 (0.64, 0.89)	0.71 (0.62, 0.82)	0.62 (0.54, 0.80)	0.70 (0.60, 0.84)
Max F_1-score median (Q1, Q3)				
PheW^{2}P2V	0.37 (0.22, 0.61)	0.20 (0.10, 0.33)	0.08 (0.04, 0.20)	0.20 (0.09, 0.38)
P2V	0.25 (0.15, 0.40)	0.08 (0.05, 0.13)	0.04 (0.02, 0.07)	0.09 (0.05, 0.20)
Regression	0.32 (0.18, 0.63)	0.18 (0.08, 0.46)	0.05 (0.02, 0.29)	0.19 (0.06, 0.46)
Random forest	0.38 (0.18, 0.66)	0.16 (0.06, 0.45)	0.05 (0.01, 0.19)	0.17 (0.05, 0.46)
Gradient boosted tree	0.37 (0.19, 0.66)	0.18 (0.07, 0.48)	0.06 (0.02, 0.30)	0.19 (0.06, 0.50)
AUC-PR median (Q1, Q3)				
PheW^{2}P2V	0.28 (0.13, 0.55)	0.10 (0.04, 0.22)	0.03 (0.01, 0.09)	0.10 (0.03, 0.27)
P2V	0.15 (0.08, 0.32)	0.03 (0.02, 0.05)	0.01 (0.00, 0.02)	0.03 (0.01, 0.11)
Regression	0.24 (0.10, 0.60)	0.09 (0.02, 0.31)	0.01 (0.00, 0.15)	0.09 (0.02, 0.35)
Random forest	0.29 (0.10, 0.63)	0.06 (0.02, 0.32)	0.01 (0.00, 0.08)	0.08 (0.01, 0.35)
Gradient boosted tree	0.29 (0.11, 0.64)	0.08 (0.02, 0.35)	0.01 (0.01, 0.17)	0.10 (0.02, 0.39)

$*$ Q1 is the 25th percentile and Q3 is the 75th percentile.

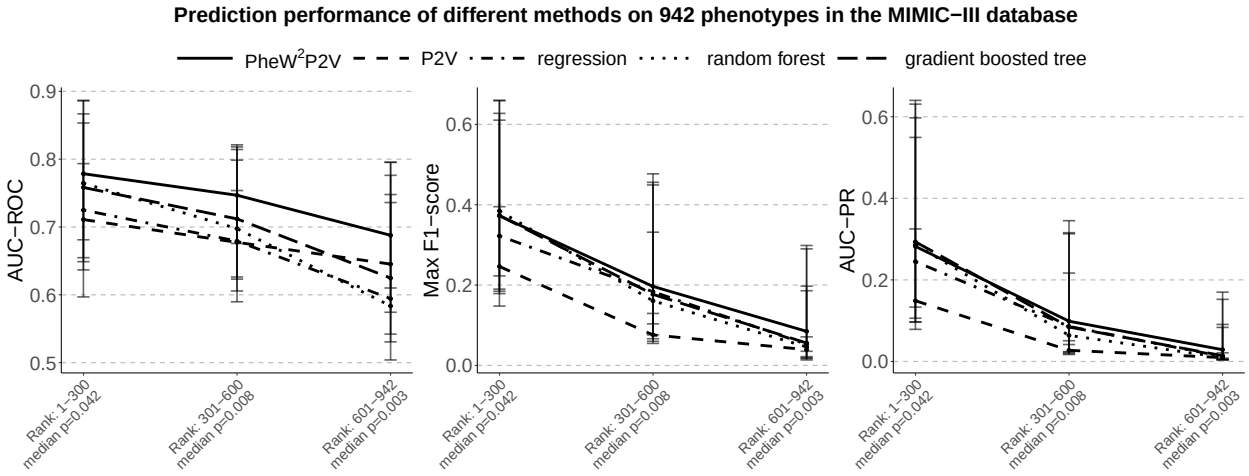

Figure 3.4: Medians, 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR across binned 300 phenotypes with descending prevalence in the MIMIC-III databases for the proposed PheW^{2}P2V, the LASSO regression, the random forest classifier, the gradient boosted tree classifier, and the unweighted P2V.

for atherosclerosis with an AUC-ROC 0.862 (competing methods have values ≤ 0.839) and a max F_1-score 0.446 (competing methods have values ≤ 0.432), suggesting that PheW^{2}P2V is better at identifying high risk patients. The second category includes rare disorders that may be diagnostically challenging, and thus may be missed if not considered in the differential diagnosis. Here, we considered adrenal hypofunction (phenotype code: 255.2), chronic pericarditis (phenotype code: 420.22), and meningitis (phenotype code: 320). PheW^{2}P2V has superior prediction performance for these conditions, for example, for adrenal hypofunction with an AUC-ROC 0.804 (competing methods have values ≤ 0.767) and a max F_1-score 0.319 (competing methods have values ≤ 0.290), suggesting that PheW^{2}P2V can be used to help with the correct diagnosis of these rare disorders. These examples demonstrate that PheW^{2}P2V is powerful in phenome-wide predictions and is capable of providing clinically-relevant data-driven risk stratification that could be useful as a screening tool to identify and flag patients with high risk conditions in early stages which may be missed overwise. Note that studies have observed that prediction tools usually have low F_1-scores to predict rare outcomes (Hunter-Zinck et al., 2019; Jeni et al., 2013; Perotte et al., 2014), which is also observed in our simulation studies with different case-control ratios summarized in Appendix B.3.

3.3.4 Computation time

We demonstrated computational efficiency of PheW^{2}P2V on a phenome-wide prediction task by comparing the computation time of PheW^{2}P2V to that of gradient boosted tree (implemented using "xgboost"), which has the second-best prediction performance among all methods (Table 3.2). PheW^{2}P2V is trained once and used for prediction across a phenome with N phenotypes with a total computation time: (training time + $N \times$ prediction time). Gradient boosted tree is trained individually for each phenotype with a total computation time for the phenome: $N*$(training time+ prediction time). With a machine of Intel(R) Xeon(R) CPU E5-2630 0 @ 2.30GHz, a phenome-wide predictions took $(213 + 8 \times N)$ seconds for PheW^{2}P2V, and $43 \times N$ seconds for gradient boosted tree. PheW^{2}P2V will be much faster when predicting a large number of phenotypes N.

Table 3.3: AUC-ROC, max F_1-score, and AUC-PR of the 5 highlighted clinical disease phenotypes in the MIMIC-III database.

Category*	Clinical disease phenotypes	Prevalence	PheW^{2}P2V	P2V	Regression	Random forest	Gradient boosted tree
	AUC-ROC						
I	Atherosclerosis	0.052	**0.862**	0.805	0.780	0.829	0.839
I	Diabetic retinopathy	0.025	**0.956**	0.910	0.864	0.932	0.934
II	Adrenal hypofunction	0.015	**0.804**	0.718	0.695	0.767	0.762
II	Chronic pericarditis	0.007	**0.818**	0.782	0.759	0.687	0.784
II	Meningitis	0.004	**0.833**	0.805	0.612	0.620	0.689
	Max F_1-score						
I	Atherosclerosis	0.052	**0.446**	0.260	0.352	0.432	0.391
I	Diabetic retinopathy	0.025	**0.521**	0.335	0.510	0.509	0.513
II	Adrenal hypofunction	0.015	**0.319**	0.088	0.214	0.290	0.267
II	Chronic pericarditis	0.007	**0.256**	0.142	0.194	0.164	0.188
II	Meningitis	0.004	**0.181**	0.119	0.047	0.035	0.050
	AUC-PR						
I	Atherosclerosis	0.052	**0.365**	0.168	0.250	0.314	0.310
I	Diabetic retinopathy	0.025	**0.462**	0.208	0.420	0.410	0.435
II	Adrenal hypofunction	0.015	**0.187**	0.043	0.121	0.161	0.151
II	Chronic pericarditis	0.007	**0.132**	0.053	0.094	0.054	0.083
II	Meningitis	0.004	**0.072**	0.046	0.010	0.008	0.013

$*$ Category I includes common conditions amenable to screening and prevention, and Category II includes rare and diagnostically challenging conditions. These categories were selected to illustrate potential clinical utility of PheW^{2}P2V predictions.

3.4 Discussion

We developed PheW^{2}P2V, a phenome-wide prediction framework that efficiently predicts phenotypes across a phenome by taking a two-step procedure, i.e., medical concept embeddings followed by tailored predictions with a novel weighting scheme. To better define phenome-wide case-control status, PheW^{2}P2V maps ICD diagnosis codes to phenotype codes. PheW^{2}P2V generates tailored patient vectors for individual phenotypes for tailored predictions. When computing patient vectors, the proposed weighting scheme upweights past medical histories that are most relevant to a phenotype of interest and thus tailors the prediction to the phenotype. The computational efficiency of phenome-wide predictions is achieved by separating embeddings and predictions, making phenome-wide predictions feasible. PheW^{2}P2V is fast, flexible, and has better prediction performance than major popular comparison methods consistently across most of the 942 phenotypes in the MIMIC-III database. PheW^{2}P2V takes advantages of the word2vec algorithm to numerically represent patients' medical concepts which avoids imputing missing concepts to convert patients' medical concepts to a sparse data matrix that is needed by most data analysis methods. We also demonstrated several clinical examples in which PheW^{2}P2V can predict less common conditions that could be diagnostically challenging or missed on a routine clinical work up, such as adrenal insufficiency or chronic pericarditis. Therefore, PheW^{2}P2V could be useful as a screening tool to prospectively identify and flag patients with high risks conditions in early stages which may be missed otherwise, such as early atherosclerosis, which is preventable with medications and lifestyle changes.

With extensive simulation studies, we demonstrated superior prediction performance of PheW^{2}P2V over four widely used comparison methods: LASSO regression, random forest, gradient boosted tree, and unweighted P2V. We also demonstrated that numeric vectors of signal medical concepts and outcome concepts are able to recover association signal strengths between them (Appendix B.2). This is promising as it suggests that all information is preserved through embedding. In contrast to traditional supervised learning methods, patients without label information (e.g., pa-

tients with only one admission in our goal of prediction) remain informative, as they can still be used to train the embedding to generate numeric representations of medical concepts. This advantage enabled us to leverage 39,001 additional patients with only one admission in the MIMIC-III database to conduct the phenome-wide predictions. Note that the transferability of the medical concept embeddings from one EHR database to another need to be studied further.

In summary, PheW^{2}P2V is the first phenome-wide prediction framework. We have demonstrated its superior prediction performance and computational efficiency using simulation studies and clinical applications on phenome-wide prediction tasks using the MIMIC-III database. Several showcases of clinical disease phenotypes suggested great potentials of PheW^{2}P2V to serve as a computable predictive tool that can aid in clinical decisions through phenome-wide predictions in a real-life clinical setting.

Chapter 4: Multi-view graph convolutional clustering with applications to cancer subtyping with multi-omics data

4.1 Introduction

High-throughput technologies in biomedical research have led to an unprecedented explosion of large-scale data. Data-driven statistical models, machine learning and deep learning techniques have played crucial roles in managing, analyzing, and interpreting these massive data, transforming them into meaningful biological insights. With the help of available multi-omics data, we can better conduct an important task in biomedical research, that is, to perform disease subtyping analysis and identify groups of subjects that are biologically similar (Wiwie et al., 2015). Multiple algorithms have been developed for disease subtyping using multi-omics data, many were applied to different cancer types to better understand the tumor heterogeneity (Mo et al., 2013; Ramazzotti et al., 2018; Rappoport & Shamir, 2018; Ruan et al., 2019; Shen et al., 2009; B. Wang, Mezlini, et al., 2014; Wei et al., 2023). Through dissecting tumor heterogeneity at molecular levels, we can identify clinically meaningful cancer subtypes that can help improve prognostics and personalized therapeutic strategies (13 et al., 2012; Curtis et al., 2012; Dagogo-Jack & Shaw, 2018; Kristensen et al., 2014; Meacham & Morrison, 2013; Yuan et al., 2014). Similar clustering algorithms have also been used for other applications, such as single-cell clustering using RNA-sequencing (RNA-seq) data to uncover cellular diversities (Kiselev et al., 2017; X. Li et al., 2020; Ranjan et al., 2021; T. Tian et al., 2019; J. Wang et al., 2021).

As an unsupervised learning problem, some traditional clustering methods include K-means (MacQueen, 1967) and Gaussian Mixture Models (Bishop & Nasrabadi, 2006) have been widely used. However, they are not effective with high dimensional multi-omics data. Many other clustering methods first learn a low dimensional latent space or a new representation of the original

data, and then apply K-means. Representation learning algorithms, such as principal component analysis (PCA) and autoencoder (one type of neural networks) (Bengio et al., 2013; Hinton & Salakhutdinov, 2006; Vincent et al., 2010), are performed on raw feature spaces; others such as spectral clustering (i.e., eigen decomposition on the graph) (Von Luxburg, 2007) and graph autoencoder (Kipf & Welling, 2016b; F. Tian et al., 2014) are performed on patient graphs (i.e., pairwise patient similarity matrices) that are constructed from high dimensional inputs. For large scale graphs, autoencoder combined with K-means is more computationally efficient than spectral clustering and is capable of feature learning with additional constrains such as sparsity penalties (F. Tian et al., 2014; Xie et al., 2016). For disease subtyping using multi-omics data, it is necessary to comprehend the interplay between different views of patient information, where each type of omics data could be considered as a view of patients (Kristensen et al., 2014; Rappoport & Shamir, 2018). To integrate multi-view/multi-omics data, a naïve approach is to concatenate all data types and perform standard clustering algorithms, which suffers from the problems of high dimensionality and imbalanced numbers of features across different views. An improved approach is to perform dimension reduction first and concatenate low dimensional representations. Many models have been developed to integrate multi-view data, from early attempts such as iCluster/iClusterPlus (Mo et al., 2013; Shen et al., 2009) [Shen et al., 2009, Mo et al., 2013] to more recent methods such as graph-based integration methods, namely similarity network fusion (SNF) (B. Wang, Mezlini, et al., 2014) and boosted similarity network fusion (abSNF) (Ruan et al., 2019), as well as kernel-based methods such as multi-kernel learning (CIMLR) (Ramazzotti et al., 2018) and hierarchical multi-kernel learning (Wei et al., 2023). These methods are usually computationally intense and not capable of conducting feature learning.

Most aforementioned methods for clustering using multi-omics data have two steps that involve feature learning to obtain an integrated latent space first and then calculate label assignments separately using K-means. To improve clustering results, deep embedded clustering (DEC) was introduced to jointly solve feature representations and cluster assignments with the latent space learned specifically for clustering (Xie et al., 2016). DEC refines a centroid-based probability dis-

tribution through leveraging samples with highly confident label assignments and simultaneously achieves feature learning and clustering in a self-learning manner.

Many existing methods for disease subtyping using integrated multi-omics data focus on patient graphs generated using similarity measures from omics data. These graphs focus on subject level aggregated omics information but ignore feature level individual molecular characteristics, which are also helpful for disease subtyping but are absent in graphs. By combining both patient graphs and molecular feature level omics data we can leverage the granularity of feature-level data and the structural insights provided by subject-level data. Similar idea has been applied in a supervised multi-omics analysis to solve classification problems for Alzheimer's disease and low-grade glioma (LGG) (T. Wang et al., 2021). Here we propose a multi-view graph convolutional clustering (MultiGCC) framework, which leverages graph convolutional encoders (Kipf & Welling, 2016a) to enhance the omics graph embeddings using molecular level features for each individual omics data. MultiGCC then simultaneously updates the enhanced graph embeddings and clustering assignments through a self-learning process to optimize a clustering objective function.

We applied the proposed MultiGCC to integrate gene expression, DNA methylation, and somatic mutation to identify subtypes of liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD) using data from The Cancer Genome Atlas (TCGA) project (https://www.cancer.gov/tcga). Cancer subtypes identified by MultiGCC are more significantly associated with patient survival than those identified by comparison methods. We further conducted analyses of molecular characteristics on the identified subtypes which provided insights of cancer heterogeneity and biological meaning of LIHC and STAD subtypes.

4.2 Methods

4.2.1 The proposed MultiGCC

We propose MultiGCC, a disease subtyping framework that integrates molecular level features into each patient omics graph constructed from individual types of omics data through graph convolutional networks. As illustrated in Figure 4.1, the MultiGCC framework has three parts: (1)

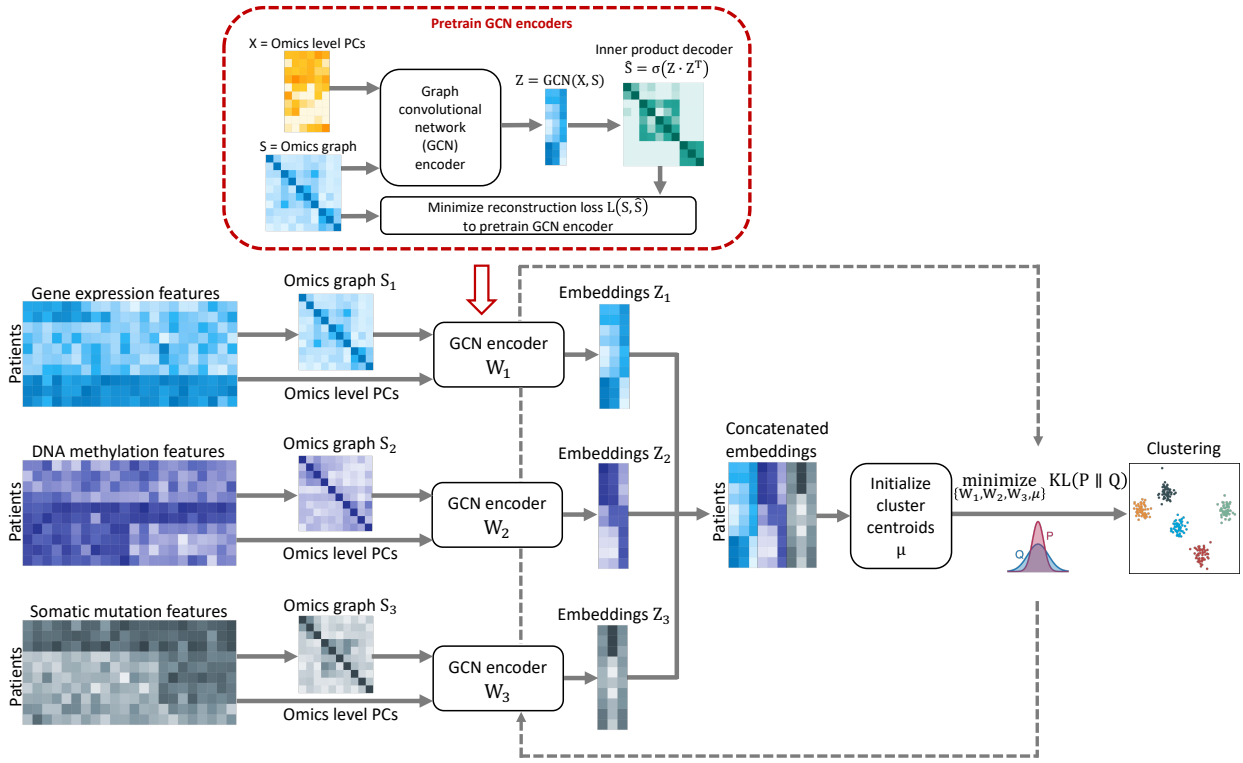

Figure 4.1: The workflow of the proposed MultiGCC framework.

constructing patient graphs from individual high dimensional omics data; (2) initializing enhanced patient graph embeddings using molecular level features through principal components (PCs); and (3) iteratively refining the enhanced patient graph embeddings and clusters assignments through a self-learning process for clustering.

Step 1. Construct patient omics graphs.

For each type of high dimensional omics data, we construct a patient graph, i.e., a similarity matrix S with a dimension $N \times N$, where N is the number of patients and each entry S_{ij} is a similarity measure between patients i and j. Suppose we have K omics data types denoted as $X^{(k)}, k = 1, 2, ..., K$, which are normalized to a scale of $[0, 1]$. We calculate pairwise cosine similarities from $X^{(k)}$ as in Equation 4.1, which was previously using in other omics analyses (T. Wang et al., 2021):

$$S_{ij}^{(k)} = \begin{cases} s\left(x_i^{(k)}, x_j^{(k)}\right), & \text{if } s\left(x_i^{(k)}, x_j^{(k)}\right) \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

Here $x_i^{(k)}$ and $x_j^{(k)}$ are the data vectors of the kth type of omics data for patients i and j, and $s\left(x_i^{(k)}, x_j^{(k)}\right) = \frac{x_i^{(k)} \cdot x_j^{(k)}}{||x_i^{(k)}||\ ||x_j^{(k)}||}$ is the cosine similarity. To construct the omics graph, we set nearest neighbors of a sample through a threshold ϵ, such that we only retain similarities for a pair of patients if the cosine similarity between them is greater than or equal to ϵ, aiming to have m neighbors per patient on average, where we usually set $m = \sqrt{N}$ with N being the total number of patients. Thus, ϵ is determined that satisfies $m = \frac{1}{N} \sum_{i,j} I\left[s\left(x_i^{(k)}, x_j^{(k)}\right) \geq \epsilon\right]$.

Step 2. Initialize enhanced embeddings of omics graphs using principal components of molecular level features.

For each omics data $X^{(k)}$, we calculate the first 50 principal components (PCs) and denote them as $C^{(k)}$. We use these PCs extracted from molecular level features to enhance the embeddings of the kth omics graph as follows:

$$\begin{aligned} Z^{(k)} &= \text{ReLU}\left(\tilde{S}^{(k)} C^{(k)} W^{(k)}\right) \\ \underset{W^{(k)}}{\text{minimize}} &\left\| S^{(k)} - \sigma\left(Z^{(k)} Z^{(k)^T}\right)\right\|^2 \end{aligned} \tag{4.2}$$

Here $\tilde{S}^{(k)} = I + S^{(k)}$, where $S^{(k)}$ is the kth omics graph, and $\tilde{S}^{(k)}$ is normalized through $\tilde{D}^{-1/2} \tilde{S}^{(k)} \tilde{D}^{-1/2}$, with $\tilde{D}$ being a diagonal matrix with $\tilde{D}_{ii} = \sum_j \tilde{S}_{ij}^{(k)}$ and I being the identity matrix to boost the self-connections in the graph (Kipf & Welling, 2016a). ReLU$(\cdot)$ and $\sigma(\cdot)$ are activation functions with ReLU$(x) = \max(0, x)$ and $\sigma(x) = 1/(1 + e^{-x})$. The encoder parameters $W^{(k)}$ are trained by minimizing the reconstruction loss of the graph $S^{(k)}$.

Step 3. Refine enhanced embeddings initialized in Step 2 through a self-learning process for clustering iteratively

The last step of MultiGCC is a self-learning process for clustering where feature level information enhanced graph embeddings and clustering assignments are iteratively refined through optimizing an objective function for clustering. Specifically, the feature level information enhanced graph embedding $Z^{(k)}$ from Step 2 are concatenated as $Z = Z^{(1)}\|Z^{(2)}\|...\|Z^{(K)}$, with which K-means is performed with different numbers of clusters to initialize a series of cluster centroids $\left\{\{\mu_r\}_{r=1}^{2}, \{\mu_r\}_{r=1}^{3}, ..., \{\mu_r\}_{r=1}^{R}\right\}$. We set $R = 6$. The final number of clusters will be chosen from $\{2, 3, ..., R\}$ after self-learning process with convergence. For each set of initialized cluster centroids $\{\mu_r\}$, we calculate a centroid-based probability distribution Q (Equation 4.3), or the soft cluster assignments q_{ir} using the Student's t-distribution as a kernel with degrees of freedom 1 (X. Li et al., 2020; Van der Maaten & Hinton, 2008) to measure the similarity between z_i (i.e., the integrated representations for patient i) and the cluster centroids μ_r:

$$q_{ir} = \frac{\left(1 + \|z_i - \mu_r\|^2\right)^{-1}}{\sum_{r'}\left(1 + \|z_i - \mu_{r'}\|^2\right)^{-1}}. \tag{4.3}$$

We refine the clusters by putting more emphasis on data points with high confidence or high q_{ir}, with an auxiliary target distribution P:

$$p_{ir} = \frac{q_{ir}^2/v_r}{\sum_{r'} q_{ir}^2/v_{r'}}, \tag{4.4}$$

where $v_r = \sum_i q_{ir}$ are soft cluster frequency of the rth cluster. We then iteratively update the encoder parameters $W^{(k)}$ and the cluster centroids $\{\mu_r\}$ through minimizing a Kullback–Leibler (KL) divergence objective function, to refine the integrated representations Z and the clustering labels obtained through q_{ir}:

$$\text{Loss} = \text{KL}(P\|Q) = \sum_i \sum_r p_{ir} \log \frac{p_{ir}}{q_{ir}}. \tag{4.5}$$

We determine the final number of clusters from 2, 3, ..., R using silhouette scores (Rousseeuw, 1987) based on the refined Z and refined clustering labels.

4.2.2 Comparison methods

We considered five comparison methods, (1) Concat-PCs: PCs of different omics data types were concatenated, and the K-means algorithm was performed to obtain the clustering labels; (2) Concat-SC: a spectral clustering method, where we concatenated spectral embeddings separately generated for each omics graph with number of eigenvectors being chosen by eigengap criteria (Ng et al., 2001). K-means was performed on the concatenated spectral embeddings with the number of clusters being chosen using silhouette scores; (3) SNF: similarity network fusion that integrates multi-omics graphs, followed by spectral clustering on the integrated graph with the number of clusters being chosen by eigengap criteria; (4) MultiGCC-noPCA: the proposed MultiGCC framework without using omics feature level PCs; and (5) MultiGCC-noKL: the proposed MultiGCC framework without the self-learning process through KL divergence.

4.3 Results

4.3.1 TCGA LIHC and STAD cancer data

We applied the proposed MultiGCC and comparison methods to identify cancer subtypes using gene expression, DNA methylation, and somatic mutation data from The Cancer Genome Atlas (TCGA) project for liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD). We used the R package TCGAbiolinks Version 2.26.0 (Colaprico et al., 2016) to obtain the multi-omics data. We conducted the same quality control (QC) steps to the two cancer types. For LIHC, there were 195 patients having gene expression data, 292 patients having DNA methylation data, and 370 patients having mutations data, resulting in 191 patients having all three types of omics data. We then removed 4 patients who did not have follow-up days information for prognosis, leading to 187 patients as our LIHC patient cohort. QC steps of STAD patients are the same and included in Appendix section C.1. For each omics data, we removed features with more than 30%

Table 4.1: Data summary of LIHC and STAD patients.

Variables		LIHC patients	STAD patients
Number of tumor samples		187	217
Number of deaths		84	74
Median survival days		1,135	1,095
Number of omics features after QC	Somatic Mutation	9,267	15,629
	Gene expression	20,107	20,266
	DNA methylation	381,569	381,478

missing. We further removed DNA methylation sites at known single nucleotide polymorphisms and corrected type I/II probe bias using the R package wateRmelon (Pidsley et al., 2013). After these two steps, we imputed missing data in gene expression and DNA methylation using K-nearest neighbor. Table 4.1 summarizes the data of LIHC and STAD patients after QC steps.

4.3.2 Overall performance of the proposed MultiGCC in LIHC and STAD

To evaluate the overall subtyping performance of the proposed MultiGCC and compare it to that of comparison methods, we conducted survival analysis using the log-rank test on the identified subtypes. Table 4.2 displays the number of clusters chosen using eigengap or silhouette scores (in parenthesis), and the number of clusters after removing clusters with sizes<5 and their corresponding survival p-values. For both LIHC and STAD, subtypes identified by the proposed MultiGCC that uses both subject- and feature-level information are most significantly associated with patient survival across all methods, while subtypes identified by comparison methods are not associated with patient survival. The four LIHC subtypes identified by MultiGCC are associated with patient survival with a p-value of 0.008. The three STAD subtypes identified by MultiGCC are associated with patient survival with a p-value of 0.014.

We also used Harrell's concordance index (C-index) (Harrell et al., 1982) to evaluate how accurate the identified subtypes can describe patient survival through a Cox model. The C-index is between 0.5 and 1, with a value of 1 representing a perfect model that always assigns higher risk scores to patients with earlier events, and a value of 0.5 representing a model that is no better

Table 4.2: Subtyping and survival analyses in two cancer types, with (1) the number of clusters chosen using eigengap or silhouette scores in parentheses, (2) number of clusters after filtering out clusters with sizes<5, and (3) corresponding survival p-values.

	LIHC		STAD	
Method	**Number of clusters**	**Survival P**	**Number of clusters**	**Survival P**
Concat-PCA	2 (2)	0.685	6 (6)	0.601
Concat-SC	4 (6)	0.307	3 (3)	0.976
SNF	3 (4)	0.115	2 (4)	0.093
MultiGCC-noPCA	4 (4)	0.251	3 (3)	0.233
MultiGCC-noKL	6 (6)	0.264	4 (4)	0.110
MultiGCC	4 (4)	**0.008**	3 (3)	**0.014**

than a coin flip. C-indexes of the subtypes identified by MultiGCC are 0.57 and 0.58 for LIHC and STAD, respectively, while C-indexes for other comparisons methods are around 0.5.

4.3.3 LIHC subtypes identified by MultiGCC

Other studies have also used TCGA omics data to identify LIHC subtypes. For example, using TCGA 183 LIHC patients, the TCGA group (Ally et al., 2017) identified three LIHC subtypes using iCluster (Shen et al., 2009) with five types of omics data, copy number variants, DNA methylation, gene expression, miRNA expression, and proteomics data. However, no significant difference was detected in overall survival (p=0.56) among the three subtypes (Ally et al., 2017). Our group previously developed a boosted SNF method, with three types of omics data (gene expression, DNA methylation, and mutations), we identified five LIHC subtypes using 161 TCGA LIHC patients which is associated with patient survival with a p-value 0.046 (Ruan et al., 2019). Since our sample size is close to the study by the TCGA consortium group, we further investigated the four subtypes identified by MultiGCC and the three clusters identified by iCluster and included comparisons in Figure 4.2. Note that among the 187 LIHC patients in our study, 17 did not have iCluster subtype labels.

Figure 4.2A displays the Kaplan-Meier survival curves of the four LIHC subtypes identified by MultiGCC. There is a clear difference in patient survival across the four MultiGCC subtypes.

For example, subtype 2 is the smallest subtype with 16 patients, having the worst survival with a median survival time 558 days. Subtype 4 has 38 patients and has the best survival with a median survival time 1791 days. These two subtypes were grouped together by iCluster and roughly formed the iCluster subtype 3, which has 61 patients.

Figure 4.2B displays the heatmap of top 500 gene expressions with smallest p-values from differential expression analysis using Kruskal-Wallis test comparing the four LIHC subtypes identified by MultiGCC. We clearly observed different patterns across the identified subtypes. For example, subtype 1 has the highest gene expression levels across the four subtypes at many Cancer Gene Census (CGC) genes (Futreal et al., 2004) such as *PDGFRB* and *LMO2*, and at many LIHC related genes such as *VIM* (L. Hu et al., 2004), *CCL21* (Shi et al., 2015), and *GIMAP1* (Huang et al., 2016). Subtype 2 has the highest gene expression levels at some important liver cancer genes such as *MAGEA6* (J.-C. Guo et al., 2019) and *GABRA3* (Y. Liu et al., 2008). Also at these two genes, subtype 2 and subtype 4 have significantly different gene expression levels with Wilcoxon test p-values 0.010 and 0.014, respectively, which might explain why MultiGCC can separate subtype 2 and subtype 4, instead of grouping them like iCluster.

Different patterns of DNA methylations across the four MultiGCC subtypes were similarly observed in Figure 4.2C, where we selected top 500 CpGs with smallest p-values from differential analysis using Kruskal-Wallis test. For example, subtype 3 has the highest methylation levels at many CpGs that located in LIHC related genes, such as cg21211053 on gene *GYS1* (G. Li et al., 2022), cg27395391 on gene *ATXN1* (Hirao et al., 2021), and cg22729438 on gene *TJP1* (X. Xu et al., 2015). Subtype 2 and subtype 4 both have higher methylation levels than other two subtypes at some CGC and LIHC related genes such as *PRDM16* (Y. Li et al., 2021) and *HYAL2* (Kim et al., 2022). In addition, subtype 4 has a significantly higher methylation level than subtype 2 at some CpGs such as cg01824625 on gene *EPHA3* with a Wilcoxon test p-value 0.003, which is also a candidate biomarker for LIHC patient prognosis (Lu et al., 2013).

Figure 4.2D displays the mutation landscape of the top 30 genes ranked by mutation frequencies. We also observed different mutation patterns across the four MultiGCC subtypes. For exam-

ple, 56% patients in subtype 2 have *TP53* mutations, but only 19% patients in subtype 3. *CTNNB1* mutations occurred in approximately 25-30% of patients in subtypes 1, 2, and 4, but only in 5% of patients in subtype 3. Mutations in these two CGC genes are considered as the cancer drivers for hepatocellular carcinoma development (Tornesello et al., 2013). In addition, mutations in another CGC and LIHC related gene *RB1* (Ahn et al., 2014) occurred in 24% of patients in subtype 3, but only in approximately 5% in subtypes 1, 2, and 4.

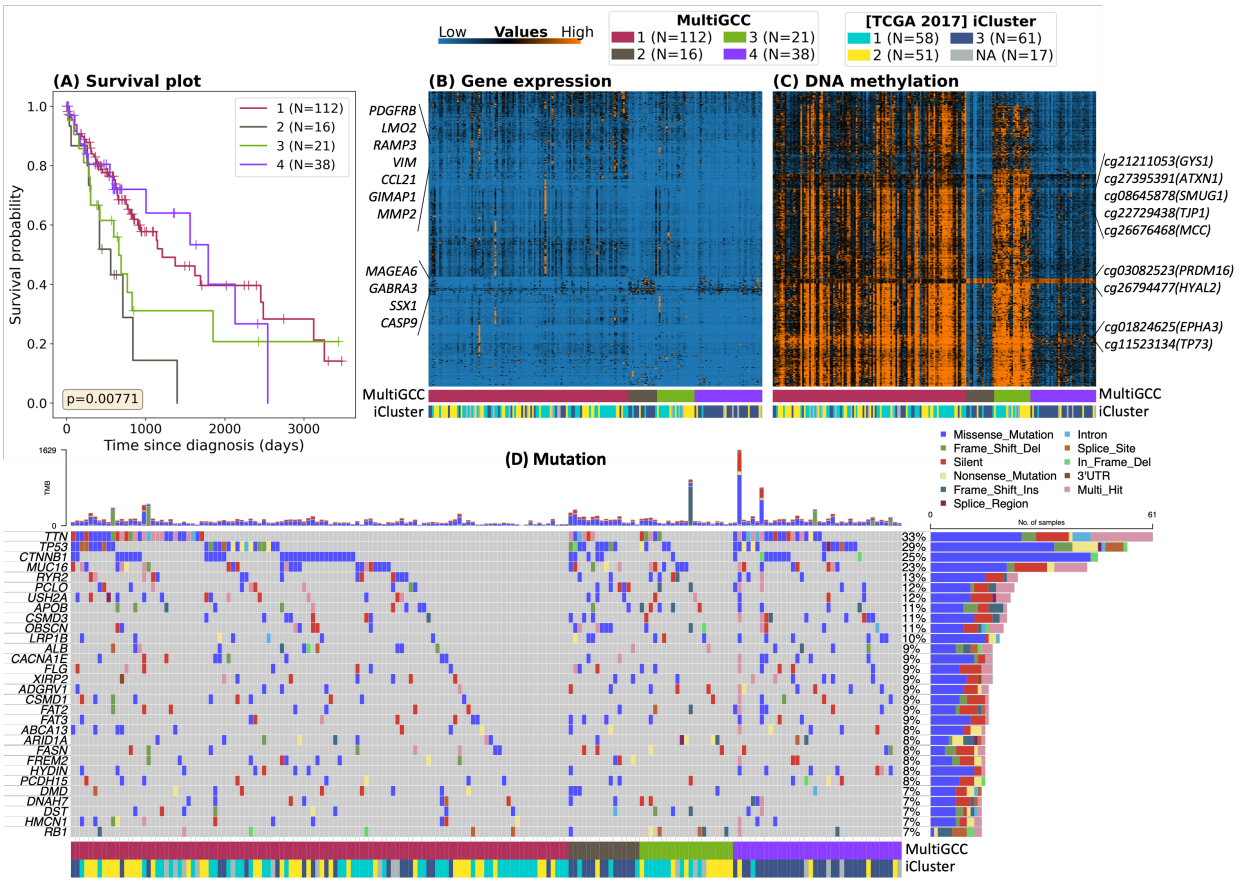

Figure 4.2: Subtyping analysis of the four LIHC subtypes identified by MultiGCC. (A) Kaplan-Meier survival curves and log-rank test p-value of the four LIHC subtypes. (B) Heatmap of top 500 gene expressions that are differentially expressed across the four LIHC subtypes by significance from the Kruskal-Wallis test. (C) Heatmap of top 500 DNA methylation CpG sites that are differentially methylated across the four LIHC subtypes by significance from the Kruskal-Wallis test. (D) The middle chart displays the heatmap of mutation profiles of the top 30 genes ranked by mutation frequencies. The top chart displays absolute number of mutation load for each sample. The right panel displays the mutation frequencies.

4.3.4 Hub gene analysis of differentially expressed genes across the four LIHC subtypes

We investigated differentially expressed genes across the four LIHC subtypes identified by MultiGCC. Using the Kruskal-Wallis test for each type of omics data separately and with a Bonferroni corrected threshold, we identified 803 differentially expressed genes, 99,002 differentially methylated CpG sites, and 2 differentially mutated genes. We selected top 200 differentially expressed genes, top 200 differentially methylated CpG sites which were mapped to 179 genes, and 2 differentially mutated genes, which leads to 381 unique genes. We investigated interactions among these 381 genes using the protein-protein interaction (PPI) network and the STRING (Search Tool for the Retrieval of Interacting Genes) database (Szklarczyk et al., 2019). Out of the 381 genes, 321 genes were mapped to the PPI network with 509 edges, where we kept edges with interaction scores ≥ 0.4, corresponding to a medium confidence that an interaction exists according to the STRING database (Szklarczyk et al., 2016). We evaluated the connectivity's of these 321 genes using three metrics: degree, stress and betweenness centrality. Degree is the number of links of a gene in a network. Stress is the total number of shortest paths from two other genes passing through a given gene. Betweenness centrality is the average frequency of a given gene appearing in the shortest paths between two other genes. These metrics measure how densely a gene is connected to or how strongly a gene is interacting with other genes in a network. We used the software Cytoscape 3.9.1 (Shannon et al., 2003) to display the network of these 321 genes (Figure 4.3) with their degrees. Genes with large degree measures are enlarged and are in the center of the network to indicate their strong interactions with other genes.

Table 4.3 lists the top five genes ranked by degree, stress, and betweenness centrality, respectively, out of the 321 genes. These top genes are known to be related to LIHC and are also differentially expressed across the four LIHC subtypes identified by MultiGCC. For example, gene *MMP2* has the highest value in degree. It is also ranked second in stress and betweenness centrality. *MMP2* gene is differentially expressed across the four subtypes with a Bonferroni corrected p-value<0.0001, with subtype 1 having the highest expression level (Figure 4.2B). It also plays a critical role in tumor invasion and metastasis and has been found to contribute to the dissemina-

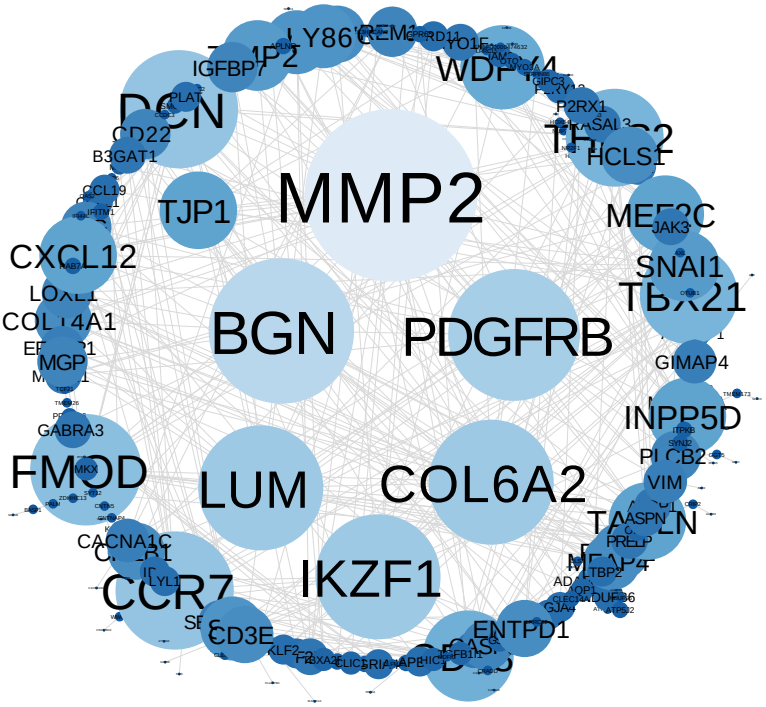

Figure 4.3: Hub gene analysis of the 321 differentially expressed genes across the four LIHC subtypes that were mapped to the PPI network. The color and size of each gene node represent the degree of each gene.

tion of liver metastases and promote the development of hepatocellular carcinoma (Musso et al., 1997; B. Wang, Ding, et al., 2014). Gene *PDGFRB* ranks the first in stress and third in degree and betweenness centrality. Gene *PDGFRB* is differentially expressed across the four subtypes with a Bonferroni corrected p-value<0.0001, where subtype 1 and subtype 3 have higher expression levels than that of other two subtypes (Figure 4.2B). *PDGFRB* gene is a CGC gene. Recent studies found that invasion of hepatocellular carcinoma cells could be inhibited through repressing *PDGFRB* expression (Luo et al., 2022; Pan et al., 2022). In addition, gene *TJP1* ranks first in betweenness centrality and third in stress. DNA methylation levels of CpG site cg22729438 in gene *TJP1* was significantly methylated across the four subtypes with a Bonferroni corrected p-value<0.0001, where subtype 3 has the highest methylation levels (Figure 4.2C). Gene *TJP1* has been found to act as a metastatic suppressor and is associated with LIHC prognosis (Nagai et al., 2016; X. Xu et al., 2015).

Table 4.3: Top five genes ranked by degree, stress, and betweenness centrality, in the PPI network of LIHC patients.

Gene	Degree	Gene	Stress	Gene	Betweenness centrality
MMP2	26	PDGFRB	21990	TJP1	0.1494
BGN	22	MMP2	21404	MMP2	0.1290
PDGFRB	20	TJP1	17340	PDGFRB	0.1250
COL6A2	19	BGN	15672	MEF2C	0.0886
LUM	19	IKZF1	15632	IKZF1	0.0870

4.3.5 STAD subtypes identified by MultiGCC

We conducted similar analyses for STAD subtypes. The three STAD subtypes identified by the proposed MultiGCC are associated with patient survival with a p-value of 0.014 (Table 4.2), while subtypes identified by comparison methods are not associated with survival. The three STAD subtypes were also studied similarly, and detailed results are included in Appendix C. The TCGA group (Network et al., 2014) also identified four subtypes using 294 STAD patients using a consensus clustering method (Wilkerson & Hayes, 2010). But these four subtypes were not significantly associated with patient survival (p=0.89).

4.4 Discussion

We developed MultiGCC, a multi-view graph convolutional clustering method for disease subtyping using multi-omics data. MultiGCC incorporates molecular feature level data into graph embeddings that are extracted from patient/graph level aggregated high dimensional multi-omics data. Using graph convolutional encoders, granular feature level information can be used to enhance graph embeddings. Through self-learning that optimizes a clustering objection function, MultiGCC simultaneously achieves feature learning and clustering assignments, where cluster separations are gradually improved by leveraging samples with high confidence in label assignments.

Graph convolutional encoders in MultiGCC are especially powerful in scenarios where feature matrices contain additional information that are not present in graphs. Patient omics graphs that

are constructed using collective information in individual types of omics data provide patient level correlations. Omics PC features provide molecular feature level characteristics. Both data are informative for disease subtyping and could generate better clustering results than methods that only use one piece of information out of two. While K-means is widely applied as the to-go clustering method in many sophisticated algorithms for clustering, it is not capable of separating clusters that are not linearly separable. Instead, MultiGCC iteratively refines cluster assignment and the latent feature space to allow nonlinear cluster separations.

In the clinical studies of two cancer types, LIHC and STAD, using gene expression, DNA methylation, and mutation data from TCGA, we demonstrated an improved clustering performance of MultiGCC over that of comparison methods. Subtypes identified by MultiGCC more accurately reflect their survival outcomes. We conducted further investigations on the biological meaning of the cancer subtypes identified by MultiGCC. Several hub genes were identified that are differentially expressed across cancer subtypes and are densely connected with many other genes in the PPI network. For example, hub genes such as *MMP2*, *PDGFRB*, and *TJP1* were identified for LIHC subtypes.

One limitation of the current study is that we treat each type of omics data equally when conduct clustering, where the model does not explicitly consider the potentially different levels of importance across different omics types. For example, some omics data types might have a very small signal-to-noise ratio for subtyping, suggesting that the clustering result might be improved if the model pays less attention to these relatively unimportant omics data types. We are currently working on extending the MultiGCC framework to dynamically consider the different importance of different omics types, such as incorporating a data-driven attention mechanism that has been widely used in natural language processing and computer vision tasks.

Chapter 5: Conclusions

In conclusion, this dissertation has made contributions in developing novel computational algorithms that use healthcare data to enhance our understanding of disease progression, improve early diagnosis of diseases, and facilitate personalized and targeted treatments. By effectively utilizing the rich information of healthcare data available in electronic health records (EHRs) and multi-omics datasets, we have developed machine learning and statistical methods that can help with advancement in biomedical research and clinical decision-making, ultimately leading to better patient outcomes.

The first approach, PsDF, demonstrated the effectiveness of a similarity-based supervised learning method for predicting specific clinical outcomes using EHR data. PsDF performed patient similarity assessment on each available domain data separately, and then integrated the affinity information over various domains into a comprehensive similarity metric. The integrated patient similarity is used to support outcome prediction by assigning a risk score to each patient. This proposed framework not only successfully predicted end stage kidney disease (ESKD) and severe aortic stenosis (AS) requiring valve replacement, but also showcased its robustness to random missingness, a common issue in real-life EHR data. By addressing the challenge of integrating multiple EHR domains, PsDF achieved more reliable predictions and improved early detection and diagnosis in various healthcare settings. One limitation of the current study is that PsDF only considers one clinical outcome at a time during the model training. Also, we use a simple imputation to code all features in EHR-based domains to be binary, indicating the presence or absence of a record, while not considering cumulative counts or continuous measures, which likely led to some information loss. Thus, we developed the second approach to overcome some of these disadvantages.

The second approach, PheW^{2}P2V, expanded the scope of our analysis to a phenome-wide

prediction task using EHR data, where we simultaneously predicted numerous phenotypes while maintaining computational efficiency and high prediction performance. PheW^{2}P2V overcomes the problems of outcome-specific and fully supervised in the traditional machine learning methods, including PsDF. PheW^{2}P2V defined clinical disease phenotypes using Phecode mapping based on ICD codes, in order to reduces redundancy and case-control misclassification in a phenome-wide prediction task. Through upweighting medical records of patients that are more relevant to a phenotype of interest in calculating patient vectors, PheW^{2}P2V achieved a tailored prediction for the phenotype. The calculation of weighted patient vectors is computationally efficient, and the weighting mechanism ensures tailored predictions across the phenome. By leveraging the Phecode mapping, numeric representations of medical concepts, and the proposed novel weighting mechanism, PheW^{2}P2V showed its superior prediction power to improve early detection and diagnosis across 942 phenotypes in the MIMIC-III database. One direction to study further and improve PheW^{2}P2V is to leverage the power of transfer learning and apply the medical concept embeddings from one EHR database to another.

The third proposed method, MultiGCC, focused on multi-omics data, integrating the multi-view patient information to identify novel disease subtypes. MultiGCC obtained an enhanced feature representation of patient information, using graph convolutional network (GCN) encoders to incorporate molecular-level features into graph embeddings of high dimensional multi-omics data. Through a self-learning process, MultiGCC jointly updated the graph embeddings and the clustering assignments to achieves a better separation of clusters. This innovative clustering framework successfully identified biologically meaningful subtypes for two cancer types, liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD). With the integration of multi-omics data, MultiGCC could help with a deeper understanding of tumor heterogeneity and improved personalized treatments. One limitation of MultiGCC is that the model do not explicitly consider the potentially different levels of importance across different omics types when conduct clustering. In our future endeavors, we intend to enhance the MultiGCC framework by dynamically considering the significance of different omics types, such as integrating a data-driven attention mechanism,

an approach that has been widely used in fields such as natural language processing and computer vision tasks.

The three methods we developed in this dissertation share a common goal of leveraging the power of data-driven approaches to provide more comprehensive perspectives of diseases, making contributions in healthcare and biomedical research. The successful application of these methods to real datasets highlights their potential to be used by clinicians as tools to provide clinical support with the help of big data. As our knowledge of diseases continues to expand, we can expect significant improvements in general patient outcomes, as well as the development of more effective personalized treatments. In summary, this dissertation has made notable contributions to the development of data-driven methods in healthcare and biomedical research, utilizing both EHR and multi-omics data - each with their own unique patient information.

The innovative methods developed in this dissertation not only contribute to existing machine learning research but also provide directions for future investigations aiming to leverage the power of big data in healthcare. As we move forward, it is crucial to continue refining these methods and developing new methods to further improve the prediction performance for supervised learning methods and clustering results for unsupervised learning methods. Building on the three proposed methods, PsDF, PheW^{2}P2V, and MultiGCC, future research can aim to extend these computational algorithms to address emerging challenges and look for new opportunities in the rapidly evolving field of precision medicine, drug discovery, drug repurposing, etc.. One critical direction is the integration of additional data sources, such as wearable devices, mobile health applications, and other patient-generated data, to augment the existing EHR and multi-omics datasets. This could lead to an even richer representation of patient information, enabling more accurate and timely predictions of clinical outcomes and response to treatments. Another direction for future research is the development of interpretable machine learning models that can provide transparent and clinically meaningful explanations for the model outputs. This would enhance the adoption of these computational algorithms among healthcare practitioners, strengthening the collaborative relationship between data-driven approaches and human expertise. Last but not least, we hope to help integrate

methodologies of machine learning methods, clinical application research, and real-world clinical practice, resulting in more generalizable and actionable insights, to ultimately improve the lives of patients worldwide.

References

13, B.

bibinitperiod W. H.

bibinitperiod H. M. S. C. L. 9. 1. P. P. J. 1. K. R., data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, G., for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, I., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*(7620), 347–355.

Agrawal, R., & Prabakaran, S. (2020). Big data in digital healthcare: Lessons learnt and recommendations for general practice. *Heredity*, *124*(4), 525–534.

Ahn, S.-M., Jang, S. J., Shim, J. H., Kim, D., Hong, S.-M., Sung, C. O., Baek, D., Haq, F., Ansari, A. A., Lee, S. Y., et al. (2014). Genomic portrait of resectable hepatocellular carcinomas: Implications of rb1 and fgf19 aberrations for patient stratification. *Hepatology*, *60*(6), 1972–1982.

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., Holt, R. A., Jones, S. J., Lee, D., Ma, Y., et al. (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, *169*(7), 1327–1341.

Anderson, A. E., Kerr, W. T., Thames, A., Li, T., Xiao, J., & Cohen, M. S. (2016). Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general united states population: A cross-sectional, unselected, retrospective study. *Journal of biomedical informatics*, *60*, 162–168.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, *28*(2), 49–60.

Arthur, D., & Vassilvitskii, S. (2006). *K-means++: The advantages of careful seeding* (tech. rep.). Stanford.

Ashfaq, A., Sant'Anna, A., Lingman, M., & Nowaczyk, S. (2019). Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, *97*, 103256.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, *10*(2-3), 191–203.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, 92–100.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

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

Chan, L., Chan, T., Cheng, L., & Mak, W. (2010). Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 467–470.

Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of general internal medicine*, *28*, 660–665.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, *8*(1), 6085.

Chen, C., Liaw, A., Breiman, L., et al. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*(1-12), 24.

Chen, X., Garcelon, N., Neuraz, A., Billot, K., Lelarge, M., Bonald, T., Garcia, H., Martin, Y., Benoit, V., Vincent, M., et al. (2019). Phenotypic similarity for rare disease: Ciliopathy diagnoses and subtyping. *Journal of Biomedical Informatics*, *100*, 103308.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor ai: Predicting clinical events via recurrent neural networks. *Machine learning for healthcare conference*, 301–318.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., et al. (2016). Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, *44*(8), e71–e71.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346–352.

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, *15*(2), 81–94.

Dai, X.-m., Zhang, Y.-h., Lin, X.-h., Huang, X.-x., Zhang, Y., Xue, C.-r., Chen, W.-n., Ye, J.-x., Lin, X.-j., & Lin, X. (2021). Sik2 represses akt/gsk3β/β-catenin signaling and suppresses gastric cancer by inhibiting autophagic degradation of protein phosphatases. *Molecular oncology*, *15*(1), 228–245.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, *6*(1), 1–25.

Del Portillo, A., Komissarova, E. V., Bokhari, A., Hills, C., de Gonzalez, A. K., Kongkarnka, S., Remotti, H. E., Sepulveda, J. L., & Sepulveda, A. R. (2019). Downregulation of friend leukemia integration 1 (fli1) follows the stepwise progression to gastric adenocarcinoma. *Oncotarget*, *10*(39), 3852.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, *31*(12), 1102–1111.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., & Crawford, D. C. (2010). Phewas: Demonstrating

the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, *26*(9), 1205–1210.

de Souza, S. M., Valiente, A. E. F., Sá, K. M., de Carvalho Juanes, C., Rodrigues, B. J., Farias, A. C. C., Campelo, C. C., de Barros Silva, P. G., & de Almeida, P. R. C. (2019). Immunoexpression of lgr4 and b-catenin in gastric cancer and normal gastric mucosa. *Asian Pacific Journal of Cancer Prevention: APJCP*, *20*(2), 519.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dharmarajan, S. H., Bragg-Gresham, J. L., Morgenstern, H., Gillespie, B. W., Li, Y., Powe, N. R., Tuot, D. S., Banerjee, T., Burrows, N. R., Rolka, D. B., et al. (2017). State-level awareness of chronic kidney disease in the us. *American journal of preventive medicine*, *53*(3), 300–307.

Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. N., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., et al. (2019). Nmr spectroscopy for metabolomics research. *Metabolites*, *9*(7), 123.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*(34), 226–231.

Fang, C., Xu, D., Su, J., Dry, J. R., & Linghu, B. (2021). Deepan: Dee p pa tient graph convolutional n etwork integrating clinico-genomic evidence to stratify lung cancers for immunotherapy. *NPJ digital medicine*, *4*(1), 14.

Farhan, W., Wang, Z., Huang, Y., Wang, S., Wang, F., Jiang, X., et al. (2016). A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR medical informatics*, *4*(4), e5977.

Fenoglio-Preiser, C., Wang, J., Stemmermann, G., & Noffsinger, A. (2003). Tp53 and gastric carcinoma: A review. *Human mutation*, *21*(3), 258–270.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R. (2004). A census of human cancer genes. *Nature reviews cancer*, *4*(3), 177–183.

Go, A. S., Chertow, G. M., Fan, D., McCulloch, C. E., & Hsu, C.-y. (2004). Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *New England Journal of Medicine*, *351*(13), 1296–1305.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naive bayes. *ICML '99: Proceedings of the sixteenth international conference on machine learning*, 258–267.

Guimaraes, D., & Hainaut, P. (2002). Tp53: A key gene in human cancer. *Biochimie*, *84*(1), 83–93.

Guo, J., Yuan, C., Shang, N., Zheng, T., Bello, N. A., Kiryluk, K., Weng, C., & Wang, S. (2021). Similarity-based health risk prediction using domain fusion and electronic health records data. *Journal of biomedical informatics*, *116*, 103711.

Guo, J.-C., Yang, Y.-J., Zhang, J.-Q., Guo, M., Xiang, L., Yu, S.-F., Ping, H., & Zhuo, L. (2019). Microrna-448 inhibits stemness maintenance and self-renewal of hepatocellular carcinoma stem cells through the magea6-mediated ampk signaling pathway. *Journal of Cellular Physiology*, *234*(12), 23461–23474.

Guo, X., Liu, X., Zhu, E., & Yin, J. (2017). Deep clustering with convolutional autoencoders. *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, 373–382.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, *18*(1), 1–15.

Henderson, J., He, H., Malin, B. A., Denny, J. C., Kho, A. N., Ghosh, J., & Ho, J. C. (2018). Phenotyping through semi-supervised tensor factorization (psst). *AMIA Annual Symposium Proceedings*, *2018*, 564.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, *103*, 103–118.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hirao, A., Sato, Y., Tanaka, H., Nishida, K., Tomonari, T., Hirata, M., Bando, M., Kida, Y., Tanaka, T., Kawaguchi, T., et al. (2021). Mir-125b-5p is involved in sorafenib resistance through ataxin-1-mediated epithelial-mesenchymal transition in hepatocellular carcinoma. *Cancers*, *13*(19), 4917.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hu, L., Lau, S. H., Tzang, C.-H., Wen, J.-M., Wang, W., Xie, D., Huang, M., Wang, Y., Wu, M.-C., Huang, J.-F., et al. (2004). Association of vimentin overexpression and hepatocellular carcinoma metastasis. *Oncogene*, *23*(1), 298–302.

Hu, S., Zhao, X., Qian, F., Jin, C., & Hou, K. (2021). Correlation between lrp1b mutations and tumor mutation burden in gastric cancer. *Computational and Mathematical Methods in Medicine*, *2021*.

Hu, Z., Melton, G. B., Arsoniadis, E. G., Wang, Y., Kwaan, M. R., & Simon, G. J. (2017). Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of biomedical informatics*, *68*, 112–120.

Huang, Z., Zhang, W., Gao, C., Ji, B., Chi, X., Zheng, W., & Wang, H. L. (2016). Dysregulation of gtpase imap family members in hepatocellular cancer. *Molecular Medicine Reports*, *14*(5), 4119–4123.

Hunter-Zinck, H. S., Peck, J. S., Strout, T. D., & Gaehde, S. A. (2019). Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. *Journal of the American Medical Informatics Association*, *26*(12), 1427–1436.

Inada, R., Sekine, S., Taniguchi, H., Tsuda, H., Katai, H., Fujiwara, T., & Kushima, R. (2015). Arid1a expression in gastric adenocarcinoma: Clinicopathological significance and correlation with dna mismatch repair status. *World Journal of Gastroenterology: WJG*, *21*(7), 2159.

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. *2013 Humaine association conference on affective computing and intelligent interaction*, 245–251.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, *3*(1), 1–9.

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT press.

Kim, S. M., Song, G. Y., Shim, A., Lee, J. H., Eom, C. B., Liu, C., Yang, Y. M., & Seki, E. (2022). Hyaluronan synthase 2, a target of mir-200c, promotes carbon tetrachloride-induced acute and chronic liver inflammation via regulation of ccl3 and ccl4. *Experimental & molecular medicine*, *54*(6), 739–752.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137–163.

Kipf, T. N., & Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N., & Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., et al. (2017). Sc3: Consensus clustering of single-cell rna-seq data. *Nature methods*, *14*(5), 483–486.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., & Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, *14*(5), 299–313.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Li, G., Zhang, P., Wei, B., Jiang, B., Li, X., Li, H., Xing, C., Wan, Y., Ma, W., & Zhou, W. (2022). Dysregulation of gys1 and gys2 correlated with poor prognosis and immune infiltrates in patients with hepatocellular carcinoma.

Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174–311ra174.

Li, L., Li, M., & Wang, X. (2020). Cancer type-dependent correlations between tp53 mutations and antitumor immunity. *DNA repair*, *88*, 102785.

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M. P., Hu, G., & Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis. *Nature communications*, *11*(1), 2338.

Li, Y., Qi, D., Zhu, B., & Ye, X. (2021). Analysis of m6a rna methylation-related genes in liver hepatocellular carcinoma and their correlation with survival. *International Journal of Molecular Sciences*, *22*(3), 1474.

Liu, Y., Li, Y.-H., Guo, F.-J., Wang, J.-J., Sun, R.-L., Hu, J.-Y., & Li, G.-C. (2008). Gamma-aminobutyric acid promotes human hepatocellular carcinoma growth through overexpressed gamma-aminobutyric acid a receptor $\alpha 3$ subunit. *World journal of gastroenterology: WJG*, *14*(47), 7175.

Liu, Z., Li, J., Hu, X., & Xu, H. (2021). Helicobacter pylori-induced protein tyrosine phosphatase receptor type c as a prognostic biomarker for gastric cancer. *Journal of Gastrointestinal Oncology*, *12*(3), 1058.

Lu, C.-Y., Yang, Z.-X., Zhou, L., Huang, Z.-Z., Zhang, H.-T., Li, J., Tao, K.-S., & Xie, B.-Z. (2013). High levels of epha3 expression are associated with high invasive capacity and poor overall survival in hepatocellular carcinoma. *Oncology Reports*, *30*(5), 2179–2186.

Luo, Z.-y., Tian, Q., Cheng, N.-m., Liu, W.-h., Yang, Y., Chen, W., Zhang, X.-z., Zheng, X.-y., Chen, M.-s., Zhuang, Q.-y., et al. (2022). Pien tze huang inhibits migration and invasion of hepatocellular carcinoma cells by repressing pdgfrb/yap/ccn2 axis activity. *Chinese Journal of Integrative Medicine*, 1–10.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

Marlin, B. M., Kale, D. C., Khemani, R. G., & Wetzel, R. C. (2012). Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, 389–398.

Meacham, C. E., & Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, *501*(7467), 328–337.

Menachemi, N., & Collum, T. H. (2011). Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, 47–55.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, *6*(1), 1–10.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, *110*(11), 4245–4250.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86–97.

Musso, O., Théret, N., Campion, J. P., Turlin, B., Milani, S., Grappone, C., & Clément, B. (1997). In situ detection of matrix metalloproteinase-2 (mmp2) and the metalloproteinase inhibitor timp2 transcripts in human primary hepatocellular carcinoma and in liver metastasis. *Journal of hepatology*, *26*(3), 593–605.

Nadkarni, G. N., Gottesman, O., Linneman, J. G., Chase, H., Berg, R. L., Farouk, S., Nadukuru, R., Lotay, V., Ellis, S., Hripcsak, G., et al. (2014). Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annual Symposium Proceedings*, *2014*, 907.

Nagai, T., Arao, T., Nishio, K., Matsumoto, K., Hagiwara, S., Sakurai, T., Minami, Y., Ida, H., Ueshima, K., Nishida, N., et al. (2016). Impact of tight junction protein zo-1 and twist expression on postoperative survival of patients with hepatocellular carcinoma. *Digestive diseases*, *34*(6), 702–707.

Network, C. G. A. R., et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, *14*.

Office of the National Coordinator for Health Information Technology. (2021). Office-based physician electronic health record adoption. *Health IT Quick-Stat #50*.

Pan, Y., Liao, X., Yang, L., Zhang, C., Wang, J., Zheng, P., Yu, G., & Song, H. (2022). Extract of marsdenia tenacissima (roxb.) moon [apocynaceae] suppresses hepatocellular carcinoma by inhibiting angiogenesis. *Frontiers in Pharmacology*, *13*, 2151.

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2014). Diagnosis code assignment: Models and evaluation metrics. *Journal of the American Medical Informatics Association*, *21*(2), 231–237.

Petitjean, A., Achatz, M., Borresen-Dale, A., Hainaut, P., & Olivier, M. (2007). Tp53 mutations in human cancers: Functional selection and impact on cancer prognosis and outcomes. *Oncogene*, *26*(15), 2157–2165.

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing illumina 450k methylation array data. *BMC genomics*, *14*(1), 1–10.

Plaut, E. (2018). From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*.

Polubriaginof, F. C., Vanguri, R., Quinnies, K., Belbin, G. M., Yahi, A., Salmasian, H., Lorberbaum, T., Nwankwo, V., Li, L., Shervey, M. M., et al. (2018). Disease heritability inferred from familial relationships reported in medical records. *Cell*, *173*(7), 1692–1704.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health information science and systems*, *2*, 1–10.

Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S., & Sidow, A. (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature communications*, *9*(1), 4453.

Ranjan, B., Sun, W., Park, J., Mishra, K., Schmidt, F., Xie, R., Alipour, F., Singhal, V., Joanito, I., Honardoost, M. A., et al. (2021). Dubstepr is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nature Communications*, *12*(1), 5849.

Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic acids research*, *46*(20), 10546–10562.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, *4*(1), 86.

Richardson, S., & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), 731–792.

Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., et al. (2011). Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS computational biology*, *7*(8), e1002141.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Ruan, P., Wang, Y., Shen, R., & Wang, S. (2019). Using association signal annotations to boost similarity network fusion. *Bioinformatics*, *35*(19), 3718–3726.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Ryu, H., Baek, S. W., Moon, J. Y., Jo, I.-S., Kim, N., & Lee, H. J. (2018). C-c motif chemokine receptors in gastric cancer. *Molecular and Clinical Oncology*, *8*(1), 3–8.

Salha, G., Hennequin, R., & Vazirgiannis, M. (2019). Keep it simple: Graph autoencoders without graph convolutional networks. *arXiv preprint arXiv:1910.00942*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498–2504.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912.

Shi, J.-Y., Yang, L.-X., Wang, Z.-C., Wang, L.-Y., Zhou, J., Wang, X.-Y., Shi, G.-M., Ding, Z.-B., Ke, A.-W., Dai, Z., et al. (2015). Cc chemokine receptor-like 1 functions as a tumour suppressor by impairing ccr7-related chemotaxis in hepatocellular carcinoma. *The Journal of pathology*, *235*(4), 546–558.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, *21*(2), 221–230.

Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one*, *13*(8), e0202344.

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, *14*, 1177932219899051.

Sun, J., Wang, F., Hu, J., & Edabollahi, S. (2012). Supervised patient similarity measure of heterogeneous patient records. *Acm Sigkdd Explorations Newsletter*, *14*(1), 16–24.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). String v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, *47*(D1), D607–D613.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., et al. (2016). The string database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, gkw937.

Tangri, N., Grams, M. E., Levey, A. S., Coresh, J., Appel, L. J., Astor, B. C., Chodick, G., Collins, A. J., Djurdjev, O., Elley, C. R., et al. (2016). Multinational assessment of accuracy of equations for predicting risk of kidney failure: A meta-analysis. *Jama*, *315*(2), 164–174.

Tian, F., Gao, B., Cui, Q., Chen, E., & Liu, T.-Y. (2014). Learning deep representations for graph clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(1).

Tian, T., Wan, J., Song, Q., & Wei, Z. (2019). Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, *1*(4), 191–198.

Tornesello, M. L., Buonaguro, L., Tatangelo, F., Botti, G., Izzo, F., & Buonaguro, F. M. (2013). Mutations in tp53, ctnnb1 and pik3ca genes in hepatocellular carcinoma associated with hepatitis b and hepatitis c virus infections. *Genomics*, *102*(2), 74–83.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(12).

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*, 395–416.

Wang, B., Ding, Y.-M., Fan, P., Wang, B., Xu, J.-H., & Wang, W.-X. (2014). Expression and significance of mmp2 and hif-1α in hepatocellular carcinoma. *Oncology letters*, *8*(2), 539–546.

Wang, B., Jiang, J., Wang, W., Zhou, Z.-H., & Tu, Z. (2012). Unsupervised metric fusion by cross diffusion. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2997–3004.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, *11*(3), 333–337.

Wang, F., Hu, J., & Sun, J. (2012). Medical prognosis based on patient similarity and expert feedback. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 1799–1802.

Wang, F., Sun, J., & Ebadollahi, S. (2011). Integrating distance metrics learned from multiple experts and its application in patient similarity assessment. *Proceedings of the 2011 SIAM International Conference on Data Mining*, 59–70.

Wang, F., Sun, J., & Ebadollahi, S. (2012). Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(1), 54–69.

Wang, F., Sun, J., Hu, J., & Ebadollahi, S. (2011). Imet: Interactive metric learning in healthcare applications. *Proceedings of the 2011 SIAM International Conference on Data Mining*, 944–955.

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., Wang, C., Fu, H., Ma, Q., & Xu, D. (2021). Scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, *12*(1), 1882.

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., & Huang, K. (2021). Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature communications*, *12*(1), 3445.

Wei, Y., Li, L., Zhao, X., Yang, H., Sa, J., Cao, H., & Cui, Y. (2023). Cancer subtyping with heterogeneous multi-omics data via hierarchical multi-kernel learning. *Briefings in Bioinformatics*, *24*(1), bbac488.

Wells, B. J., Chagin, K. M., Nowacki, A. S., & Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *Egems*, *1*(3).

Wilkerson, M. D., & Hayes, D. N. (2010). Consensusclusterplus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics*, *26*(12), 1572–1573.

Wiwie, C., Baumbach, J., & Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature methods*, *12*(11), 1033–1038.

Wu, J., Roy, J., & Stewart, W. F. (2010). Prediction modeling using ehr data: Challenges, strategies, and a comparison of machine learning approaches. *Medical care*, S106–S113.

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S.-S. (2014). Forestexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, *67*, 105–116.

Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. *International conference on machine learning*, 478–487.

Xu, J., Murphy, S., Kochanek, K., Bastian, B., & Arias, E. (2016). Deaths: Final data for 2013 national vital statistics reports; vol 64 no 2. *Hyattsville, MD: National Center for Health Statistics*.

Xu, X., Liu, Z., Zhou, L., Xie, H., Cheng, J., Ling, Q., Wang, J., Guo, H., Wei, X., & Zheng, S. (2015). Characterization of genome-wide tfcp2 targets in hepatocellular carcinoma: Implication of targets fn1 and tjp1 in metastasis. *Journal of Experimental & Clinical Cancer Research*, *34*(1), 1–11.

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L. A., Xu, Y., Hess, K. R., Diao, L., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology*, *32*(7), 644–652.

Zein, J. G., Wu, C.-P., Attaway, A. H., Zhang, P., & Nazha, A. (2021). Novel machine learning can predict acute asthma exacerbation. *Chest*, *159*(5), 1747–1757.

Zhang, J., Kowsari, K., Harrison, J. H., Lobo, J. M., & Barnes, L. E. (2018). Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, *6*, 65333–65346.

Zhang, P., Wang, F., Hu, J., & Sorrentino, R. (2014). Towards personalized medicine: Leveraging patient similarity and drug similarity analytics. *AMIA Summits on Translational Science Proceedings*, *2014*, 132.

Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., & Wang, F. (2016). Measuring patient similarities via a deep architecture with medical concept embedding. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 749–758.

Appendix A: Appendix to similarity-based health risk prediction using domain fusion and electronic health records data

A.1 Results of ESKD prediction tools with 1:1 case/control ratio

Figure A.1 summarizes prediction performance of the four methods for the two inclusion criteria used, with 1:1 case/control ratio. The patterns are very similar to that with 1:5 case/control ratio (Figure 2.5 in the main text).

As expected, with the 1:1 case/control ratio (Figure A.1), all four methods have higher F_1-scores, F_2-scores and recalls than those with 1:5 case/control ratio (Figure 2.5 in the main text). AUCs, F_1-scores and F_2-scores are all relatively more stable for the PsDF method than other competing methods, demonstrating a clear advantage of PsDF. In addition, F_1-scores and F_2-scores of random forest decrease slower than those with 1:5 case/control ratio. This is consistent with what we observed in simulation results.

A.2 ESKD prediction tools with an additional domain of geocoding

As there are limited geocoding information available for some of the EHR patients, for demonstration purposes that PsDF can fuse all available domains, we repeated the construction of the ESKD prediction tools including the geocoding domain. We updated the samples selection for the training and test sets accordingly. There are two continuous variables available for the geocoding domain, median household income in dollars and distance to the nearest major road in meters. Other five domains are the same as described in the main text.

Hence, the comprehensive patient data included 6 domains: 1) laboratory tests, 2) ICD based diagnosis history, 3) drug exposures, 4) medical procedures, 5) demographic information with gender and race (white vs. non-white), and 6) geocoding data with income and distance to major

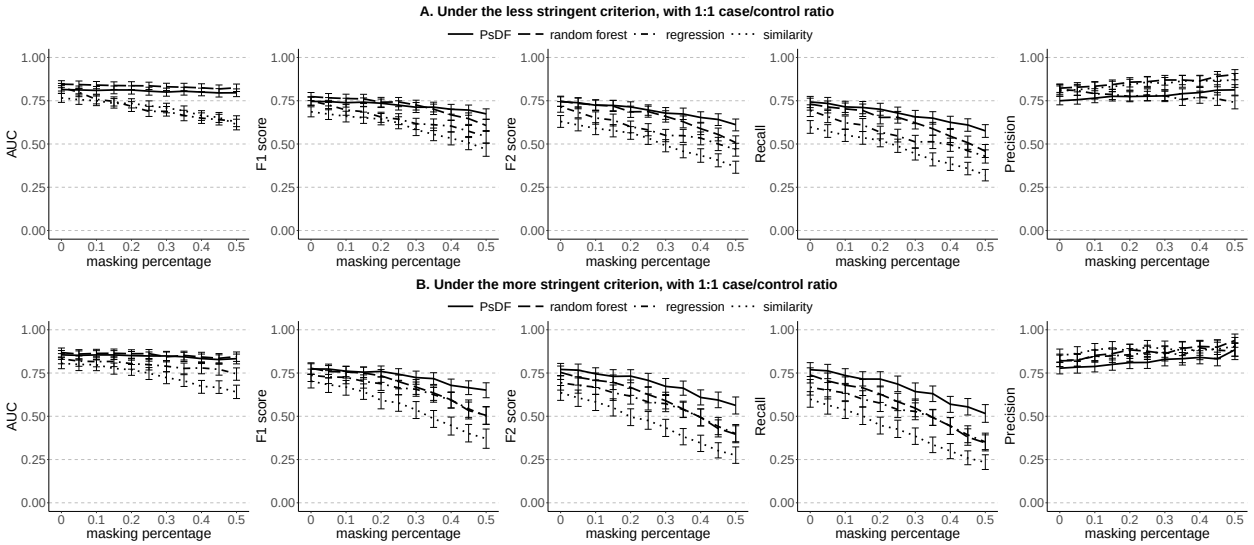

Figure A.1: With the 1:1 case/control ratio, prediction performance of the ESKD prediction tools built by the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method when the masking percentage p_{mask} increases, under two different inclusion criteria: A) the less stringent criterion, and B) the more stringent criterion.

roads. We used two different inclusion criteria to define eligible patients, a less stringent criterion that only requires patients to have demographic and geocoding domain; and a more stringent criterion that requires patients to have demographic and geocoding domain as well as records across all four EHR domains.

EHR data preprocessing for ESKD prediction

As described in the main text Section 2.3.2, we considered those 2,080 patients who were non-ESKD in 2006 but reached ESKD before 2016 as our incident ESKD cases, and those 353,295 non-ESKD patients who remained non-ESKD between 2006 and 2016 as our controls.

After requiring all patients to have demographic and geocoding data, we had 1,884 ESKD cases and 306,222 non-ESKD controls. We then applied two different inclusion criteria on the four EHR domains to define eligible patients in the study: 1) the less stringent inclusion criterion which does not have any requirement on EHR domains; 2) the more stringent inclusion criterion which requires patients to have records across all four EHR domains. Figure A.2 displays the data preprocessing pipeline and the final sample sizes with the two inclusion criteria.

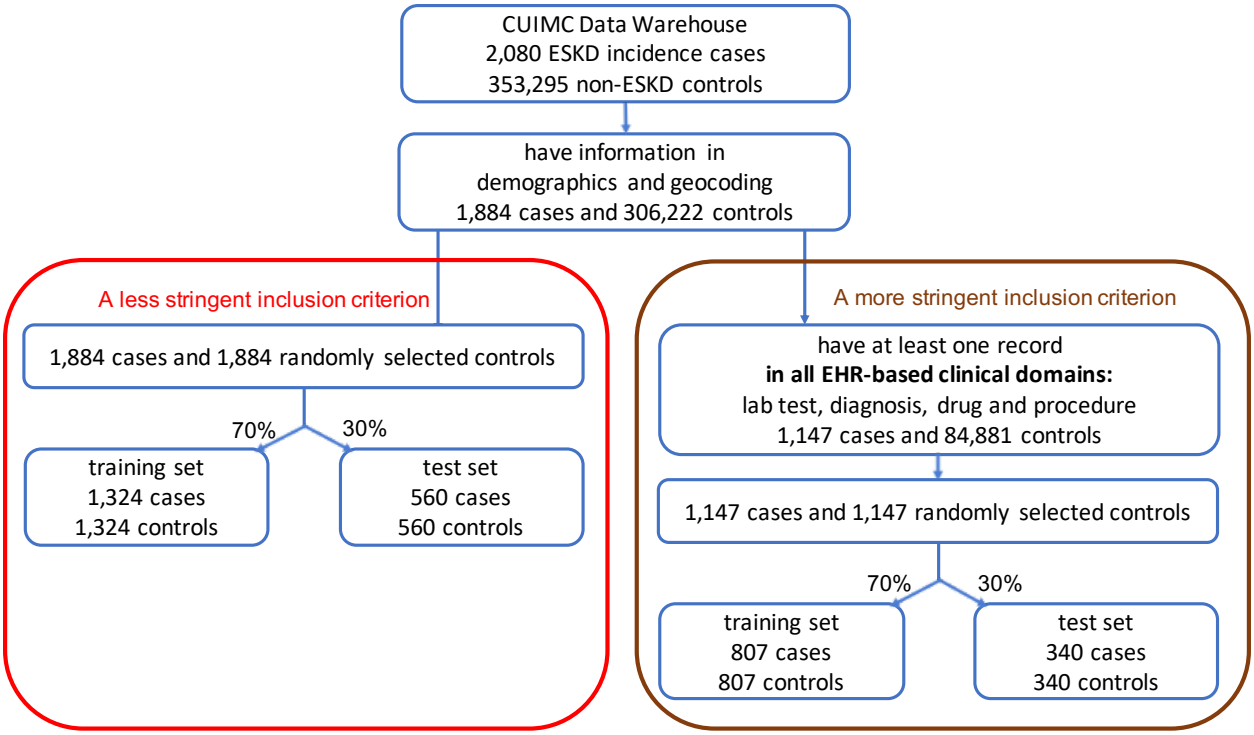

Figure A.2: ESKD data preprocessing pipeline with an additional domain of geocoding, with two different inclusion criteria to define eligible patients.

A less stringent inclusion criterion

Patients were included if they had demographic and geocoding information, resulting in 1,884 ESKD patients and 306,222 non-ESKD controls. We then randomly selected 1,884 patients among 306,222 non-ESKD controls to create a balanced case control design. We split 1,884 ESKD cases and 1,884 non-ESKD controls into two cohorts, one as the training set with 1,324 ESKD cases and 1,324 non-ESKD controls, the other as the test set with 560 ESKD cases and 560 non-ESKD controls to test the prediction performance of PsDF and the three competing methods.

A more stringent inclusion criterion

Patients were included if they had demographic and geocoding information as well as records in all four EHR domains, resulting in 1,147 ESKD patients and 84,881 non-ESKD controls. We then randomly selected 1,147 patients among the 84,881 non-ESKD controls to make a balanced case control design. We similarly split 1,147 ESKD cases and 1,147 non-ESKD controls into two cohorts, one as the training set with 807 ESKD cases and 807 non-ESKD controls, the other as the

test set with 340 ESKD cases and 340 non-ESKD controls.

In order to investigate the method's performance under an unbalanced case-control study, we parallelly did another study with 1:5 case/control ratio. In addition to the already selected controls, we further randomly selected 7,536 controls (four times of 1,884 cases) for the less stringent inclusion criterion, and 4,588 controls (four times of 1,147 cases) for the more stringent inclusion criterion. For each criterion, we split these additional control samples into two groups with ratio 70% and 30%, then added them into the training set and test set accordingly.

Feature selection using LASSO regression and random forest

Similar to the main text Section 2.3.2, we included a screening step to pre-select potentially relevant features using LASSO regression and random forest in order to capture both linear and nonlinear features for prediction, on the training set with 1:1 case/control ratio.

For stability selection using LASSO regression, among 1000 times repeated subsampling, predictors with selection probability greater than 0.6 were selected. With the training set defined by the less stringent inclusion criterion, we selected 19 features out of 1,123 laboratory tests, 30 of 7,980 diagnostic history features, 23 of 3,936 drug exposure features, 26 of 6,324 medical procedure features, gender out of the two demographic variables, and distance to major road out of the two geocoding variables. With the training set defined by the more stringent inclusion criterion, we selected 19 laboratory tests, 22 diagnostic history, 26 drug exposures, 19 medical procedures, gender and race, and distance to major road.

For feature selection using random forest, we selected features with high importance, which is defined as mean decrease accuracy. We used a threshold of greater than 0.1. With the training set defined by the less stringent inclusion criterion, we selected 24 laboratory tests, 57 diagnostic history, 46 drug exposures, 40 medical procedures, gender, and income. With the training set defined by the more stringent inclusion criterion, we selected 17 laboratory tests, 25 diagnostic history, 26 drug exposures, 27 medical procedures, and gender.

We unionized the features selected by LASSO and random forest, which led to 193 features in

total for the dataset of less stringent inclusion criterion and 121 features in total for the dataset of more stringent inclusion criterion.

Comparison of the four prediction methods

To compare the prediction performance of the four methods using the test set, we applied bootstrapping 1,000 times on the test set and obtained average AUCs, F_1-scores, F_2-scores, recalls and precisions when the threshold for the probability of being a case is set at 0.5, as well as their 95% CIs. In addition, we conducted a sensitivity analysis where we masked certain percentages of observations to investigate the robustness of PsDF and the three competing methods to missing data, as previously explained. Figure A.3 and Figure A.4 summarize prediction performance for the two inclusion criteria used, with 1:1 and 1:5 case/control ratios, respectively.

In general, the patterns are very similar to those with 5 domains. Both PsDF and random forest outperform logistic regression and the naïve similarity method in terms of AUCs (Figure A.3 and Figure A.4). AUCs drop dramatically for the logistic regression and the naïve similarity methods, while F_1-scores and F_2-scores drop quickly for random forest, with increasing masking percentage. In contrast, AUCs, F_1-scores and F_2-scores are all relatively stable for the PsDF method. This clinical study with an additional geocoding domain further demonstrates that PsDF had a more robust prediction preformation than other three methods, in the presence of random missingness in EHR data.

In addition, we observed similar AUCs, F_1-scores and F_2-scores when compared to the ESKD prediction study without geocoding (Figure 2.5 in the main text and Figure A.1), potentially indicating that the geocoding information has limited predictability in this ESKD prediction task.

A.3 Aortic Stenosis (AS) prediction tools

Among all patients in the Columbia University Irving Medical Center's EHR data warehouse as of year 2006, 5,400,082 patients without AS were defined as controls. Among these patients, as of year 2016, 6,300 developed incident severe AS requiring valve replacement between 2006

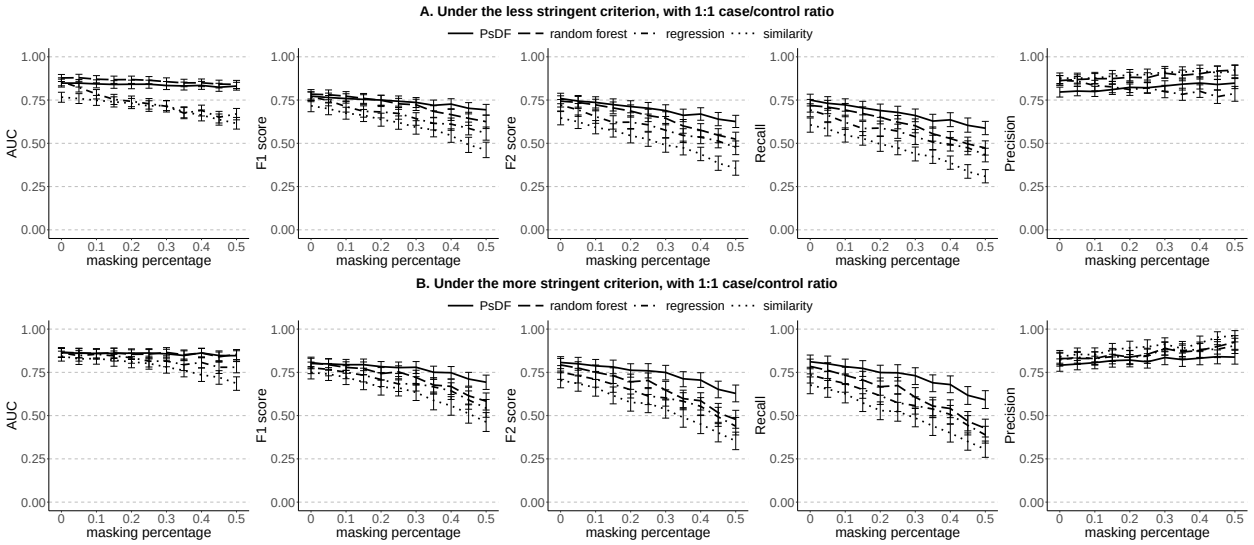

Figure A.3: With 1:1 case/control ratio, prediction performance of the ESKD prediction tools with an additional domain of geocoding, built by the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method when the masking percentage p_{mask} increases, under two different inclusion criteria: A) the less stringent criterion, and B) the more stringent criterion.

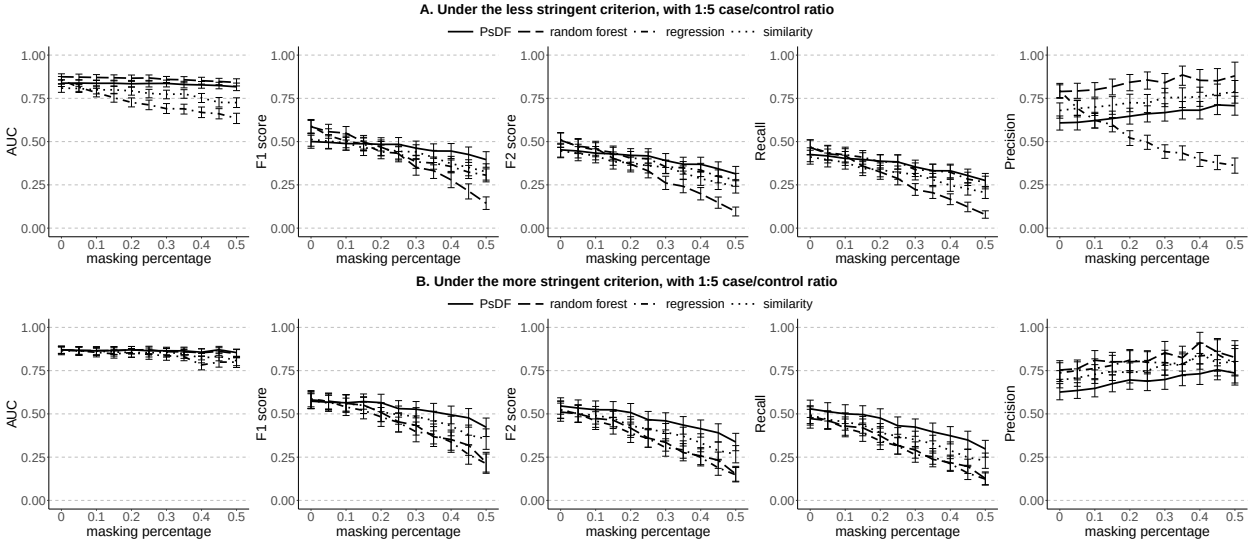

Figure A.4: With 1:5 case/control ratio, prediction performance of the ESKD prediction tools with an additional domain of geocoding, built by the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method when the masking percentage p_{mask} increases, under two different inclusion criteria: A) the less stringent criterion, and B) the more stringent criterion.

and 2016, 5,393,782 remained controls. We considered the 6,300 patients without AS in 2006 but had valve replacement before 2016 as our incident AS case samples. Due to the large sample size of control group, we randomly extracted 18,900 (three times of 6,300 AS cases) among those 5,393,782 patients who were controls in 2006 as well as in 2016, as our control samples. Our data processing pipeline is summarized in Figure A.5. The comprehensive patient data included: 1) laboratory tests, 2) ICD based diagnosis history, 3) drug exposures, 4) medical procedures, and 5) demographic information with gender and race (white vs. non-white).

We applied the PsDF algorithm and the three competing methods to build AS prediction tools and compared their prediction performance. We predicted incident AS cases between 2006-2016 using comprehensive EHR data collected in 2006 and prior. We used two different inclusion criteria to define eligible patients in the study, a less stringent inclusion criterion that requires patients to have demographic information and records in at least one of the four EHR domains; and a more stringent inclusion criterion that requires patients to have demographic information and records across all four EHR domains. Random masking procedure was also conducted at 5% to 50% with an increment of 5% in the testing set.

EHR data preprocessing for AS prediction

After requiring all patients to have information in demographics, we had 6,300 AS cases and 18,900 controls. We then applied two different inclusion criteria on the four EHR domains to define eligible patients in the study: 1) the less stringent inclusion criterion which requires patients to have records in at least one of the four EHR domains; 2) the more stringent inclusion criterion which requires patients to have records across all four EHR domains. Noted that the less stringent inclusion criterion is different from that in ESKD showcase for which there is no requirement on EHR domains, because among 6,300 AS cases and 18,900 controls, more than 70% patients do not any records in EHR domains. Figure A.5 displays the data preprocessing pipeline and the final sample sizes with the two inclusion criteria.

A less stringent inclusion criterion

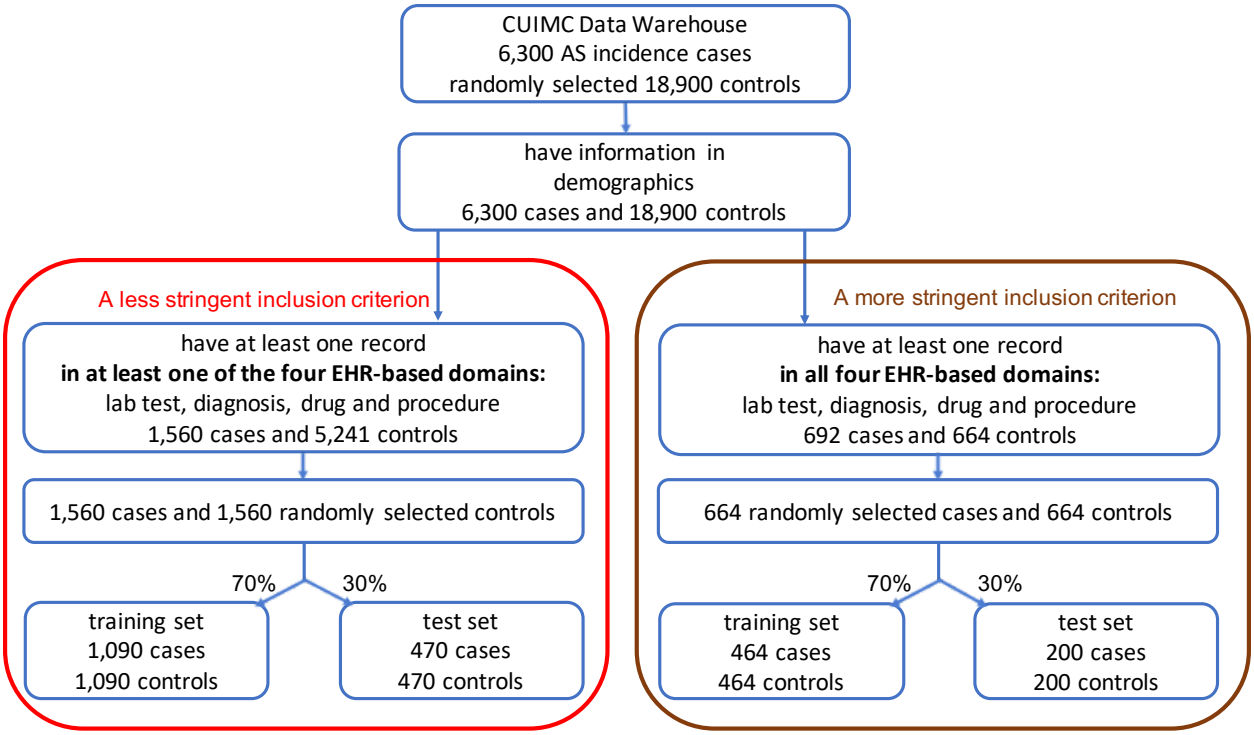

Figure A.5: AS data preprocessing pipeline with two different inclusion criteria to define eligible patients.

Patients were included if they had demographic information and records in at least one of the four EHR domains, resulting in 1,560 AS patients and 5,241 controls. We then randomly selected 1,560 patients among 5,241 controls to make a balanced case control design. We split 1,560 AS cases and 1,560 controls into two cohorts, one as the training set with 1,090 AS cases and 1,090 controls, the other as the test set with 470 AS cases and 470 controls to test the prediction performance of PsDF and the two competing methods.

A more stringent inclusion criterion

Patients were included if they had demographic and records across all four EHR-based domains, resulting in 692 AS patients and 664 controls. We then randomly selected 664 patients among 692 AS patients to make a balanced case control design. We similarly split 664 AS cases and 664 controls into two cohorts, one as the training set with 464 AS cases and 464 controls, the other as the test set with 200 AS cases and 200 controls.

Feature selection using LASSO regression and random forest

Similar to the ESKD showcase, we included a screening step to pre-select potentially relevant features using LASSO regression and random forest in order to capture both linear and nonlinear features for prediction, before applying PsDF, logistic regression and naïve similarity method.

For stability selection using LASSO regression, among 1000 times repeated subsampling, predictors with selection probability greater than 0.6 were selected. With the training set defined by the less stringent inclusion criterion, we selected 20 out of 865 laboratory tests, 42 out of 3,486 diagnostic history features, 16 out of 1,853 drug exposure features, 51 out of 1,928 medical procedure features, gender and race out of demographic information. With the training set defined by the more stringent inclusion criterion, we selected 9 laboratory tests, 15 diagnostic history, 12 drug exposure, 14 medical procedures, and gender.

For feature selection using random forest, we selected features with high importance, which is defined as mean decrease accuracy. We used a threshold of greater than 0.1. With the training set defined by the less stringent inclusion criterion, we selected 22 laboratory tests, 72 diagnostic history, 17 drug exposures, 51 medical procedures, gender and race. With the training set defined by the more stringent inclusion criterion, we selected 5 laboratory tests, 17 diagnostic history, 18 drug exposures, 19 medical procedures, and gender.

We unionized the features selected by LASSO and random forest, which led to 216 features in total for the dataset of less stringent inclusion criterion and 77 features in total for the dataset of more stringent inclusion criterion.

Comparison of the four prediction methods

To compare the prediction performance of the four methods using the test set, we applied bootstrapping 1,000 times on the test set and obtained average AUCs, F_1-scores, F_2-scores, recalls and precisions when the threshold for the probability of being a case is set at 0.5, as well as their 95% CIs. In addition, we conducted a sensitivity analysis where we masked certain percentages of observations to investigate the robustness of PsDF and the three competing methods to missing

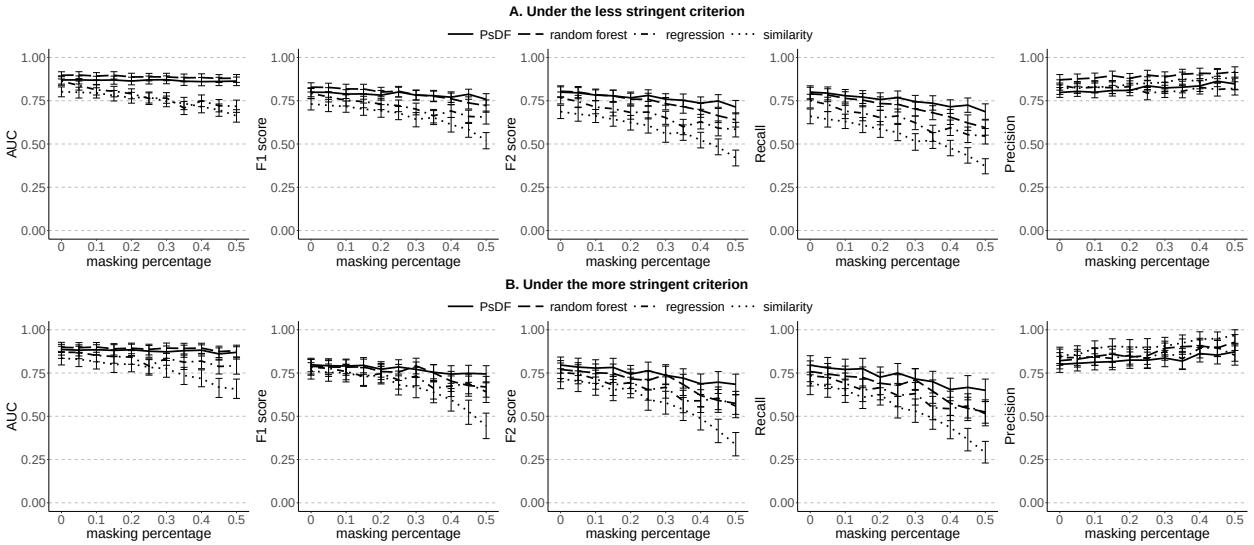

Figure A.6: Prediction performance of the AS prediction tools built by the PsDF algorithm, the random forest classifier, the logistic regression and the naïve similarity method when increasing the masking percentage p_{mask}, under two different inclusion criteria: A. the less stringent criterion, and B. the more stringent criterion.

data, as previously explained. Figure A.6 summarizes prediction performance for the two inclusion criteria used.

In general, both PsDF and random forest outperform logistic regression and the naïve similarity method in terms of AUCs (Figure A.6). With the same pattern observed in the ESKD prediction study, prediction performance (AUCs, F_1-scores and F_2-scores) of random forest, logistic regression and naïve similarity method dropped quickly when p_{mask} was increasing, but those of PsDF are relatively stable. It further indicates that PsDF had a more robust prediction preformation, in the presence of random missingness in EHR data.

The patterns of recalls and precisions are also similar to those in the ESKD prediction study and those in simulation scenario when cases have more EHR records than controls. The recalls of PsDF are higher and also drop slower than other three methods, while the precisions of PsDF are relatively lower (Figure A.6), because generally AS cases may have more observed EHR records than controls.

Appendix B: Appendix to PheW^{2}P2V - a phenome-wide prediction framework with weighted patient representations using electronic health records

B.1 Details of simulation settings

We conducted simulation studies to evaluate the prediction performance of PheW^{2}P2V and that of comparison methods. In specific, we simulated a population pool of 20,000 patients each with a binary phenotype concept C_0 and 150 unique binary medical concepts, including 10 signal predictor concepts (denoted as $C_1, C_2, \ldots, C_{10}$) that predict C_0 and 140 noise concepts (denoted as $C_{11}, C_{12}, \ldots, C_{150}$). Those 150 binary concepts were generated to be correlated with each other, to mimic potential correlations between medical concepts. To do so, for each patient, we first generated 150 continuous concepts $\widetilde{C_1}, ..., \widetilde{C_{150}}$ from a multivariate normal distribution with mean 0 and a covariance matrix, where we set the 10 signal concepts to be correlated with $\rho_{ij} = 0.6$, the 140 noise concepts to be correlated with $\rho_{ij} = 0.05$, and $\rho_{ij} = 0.05$ between signal and noise concepts. We then applied the "nearestSPD" algorithm (Higham, 1988) to find the nearest positive-definite matrix as the covariance matrix. After obtaining 150 continuous concepts, to mimic real EHR patient medical concept sequences, we median-dichotomized them into binary concepts with 1 indicating a medical concept is recorded. Then, the outcome phenotype concept C_0 was generated using a logistic model with the 10 signal predictor concepts. Other steps are included in the main text. Figure B.1 displays the steps of simulation studies.

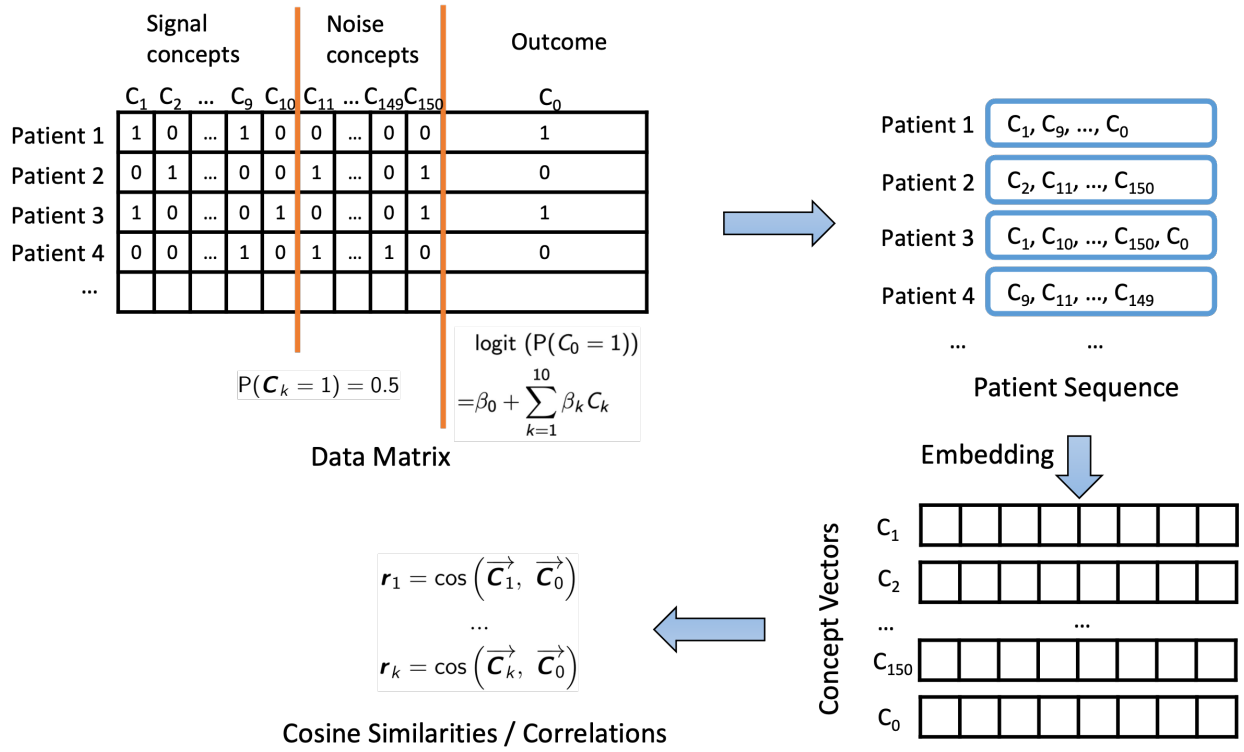

Figure B.1: Steps of simulation studies for PheW^{2}P2V.

B.2 Numeric representations can recover the association strength

With the same simulation settings, we investigated whether numeric representations of medical concepts using word2vec are able to recover the association strength between a medical concept (i.e., an explanatory predictor) and a phenotype (i.e., an outcome). Given a training set with 100 cases and 100 controls, we performed word2vec to obtain numeric representations for all 151 concepts (10 signal predictor concepts, 140 noise concepts, and 1 outcome concept). We then calculated cosine similarities between the vector of outcome concept $\overrightarrow{C_0}$ and vectors of signal/noise concepts $\overrightarrow{C_k}(k = 1, ..., 150)$, and obtained their medians, 25th and 75th percentiles across 1,000 training sets.

Figure B.2 displays the results when association signal strengths range β from -1.0 to 1.0. We can see that as the effect size increases, i.e., when absolute values of β for signal concepts in logistic regressions increase, absolute values of cosine similarities between numeric vectors of

signal concepts and the outcome concept also increase (Figure B.2A), while cosine similarities between vectors of noise concepts and that of the outcome concept are close to 0 (Figure B.2B). These results demonstrate that numeric representations of medical concepts preserve the original association signal strength nicely.

It is worth noting that a medical concept that is positively associated with an outcome in a logistic regression model (i.e., with a positive β coefficient) is also positively correlated with the outcome when being evaluated using numerically represented vectors (i.e., with a positive cosine similarity). This is expected, because according to word2vec, a positive cosine similarity will be observed between two numeric vectors for two concepts (e.g., C_1 and C_0) when they have similar contexts, i.e., when nearby concepts of C_1 and nearby concepts of C_0 are similar. When the signal concept C_1 is positively associated with the outcome concept C_0, the probability of having outcome C_0 when concept C_1 exists is high, i.e., C_1 and C_0 usually appear simultaneously in a patient's medical concept sequence. As a result, C_1 and C_0 will have similar contexts which result in a positive correlation between their numeric vectors. Similar explanation goes to scenarios when C_1 and C_0 are negatively associated with a negative β coefficient, their numeric vectors will be negatively correlated. We thus used positive β coefficients in the following simulation studies to evaluate prediction performance of PheW^{2}P2V.

B.3 Simulation studies with more case/control ratios

In addition to the 1:19 case-control ratios in the main text, we also considered unbalanced simulation scenarios ranging case/control ratios from 1:1, 3:7, 1:9 to 1:19, with other simulation settings unchanged. We summarized medians, 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR across 1,000 test sets in Figure B.3. We observed that the improvement of PheW^{2}P2V over LASSO regression, random forest, and gradient boosted tree increases as the case-control design becomes more unbalanced. This is because the imbalance affects the prediction performance of regression-based models like LASSO and tree-based models more.

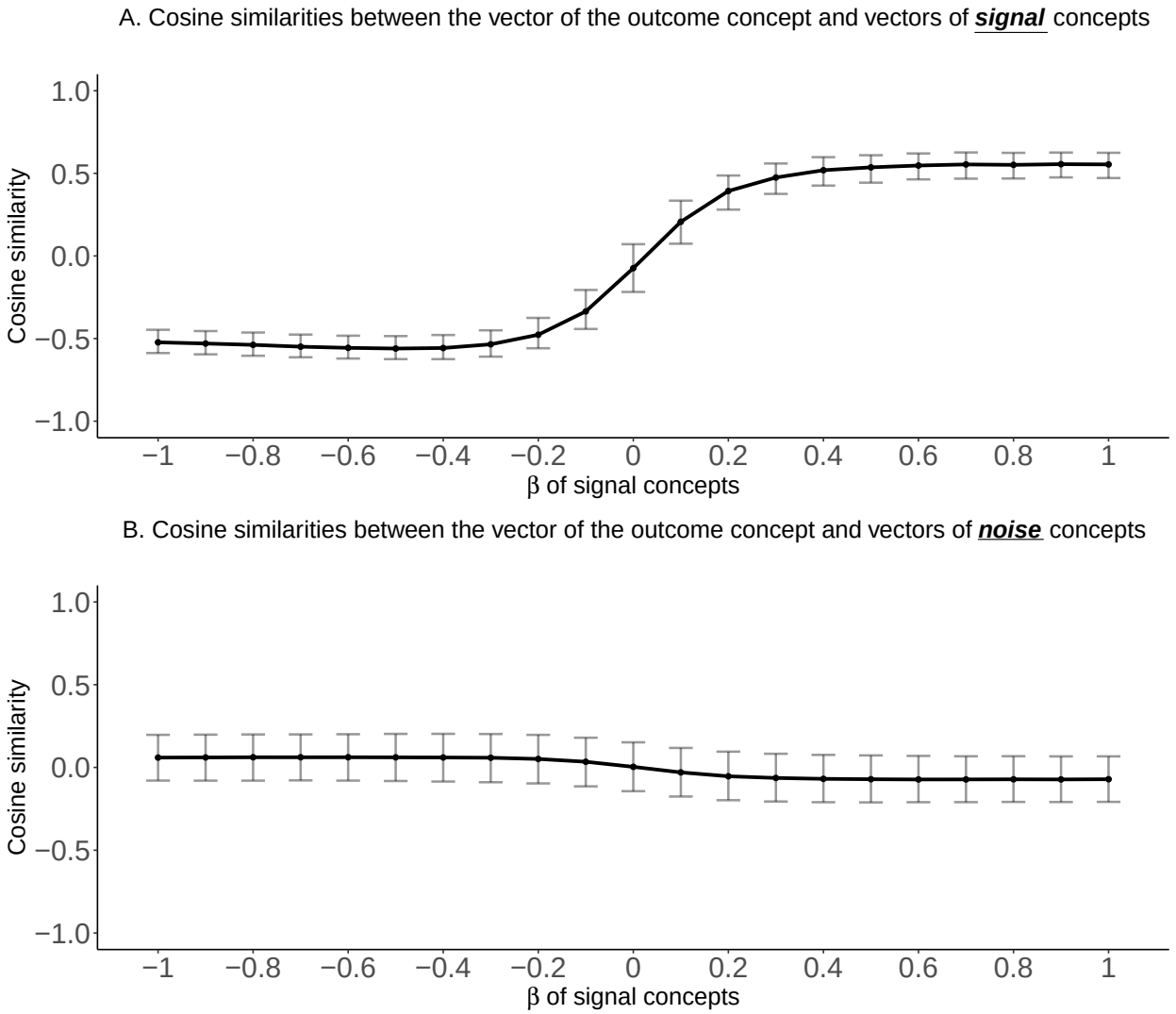

Figure B.2: Simulation results of medians and 25th and 75th percentiles of cosine similarities between vectors of 10 signal concepts (A), vectors of 140 noise concepts (B) and vector of the outcome concept.

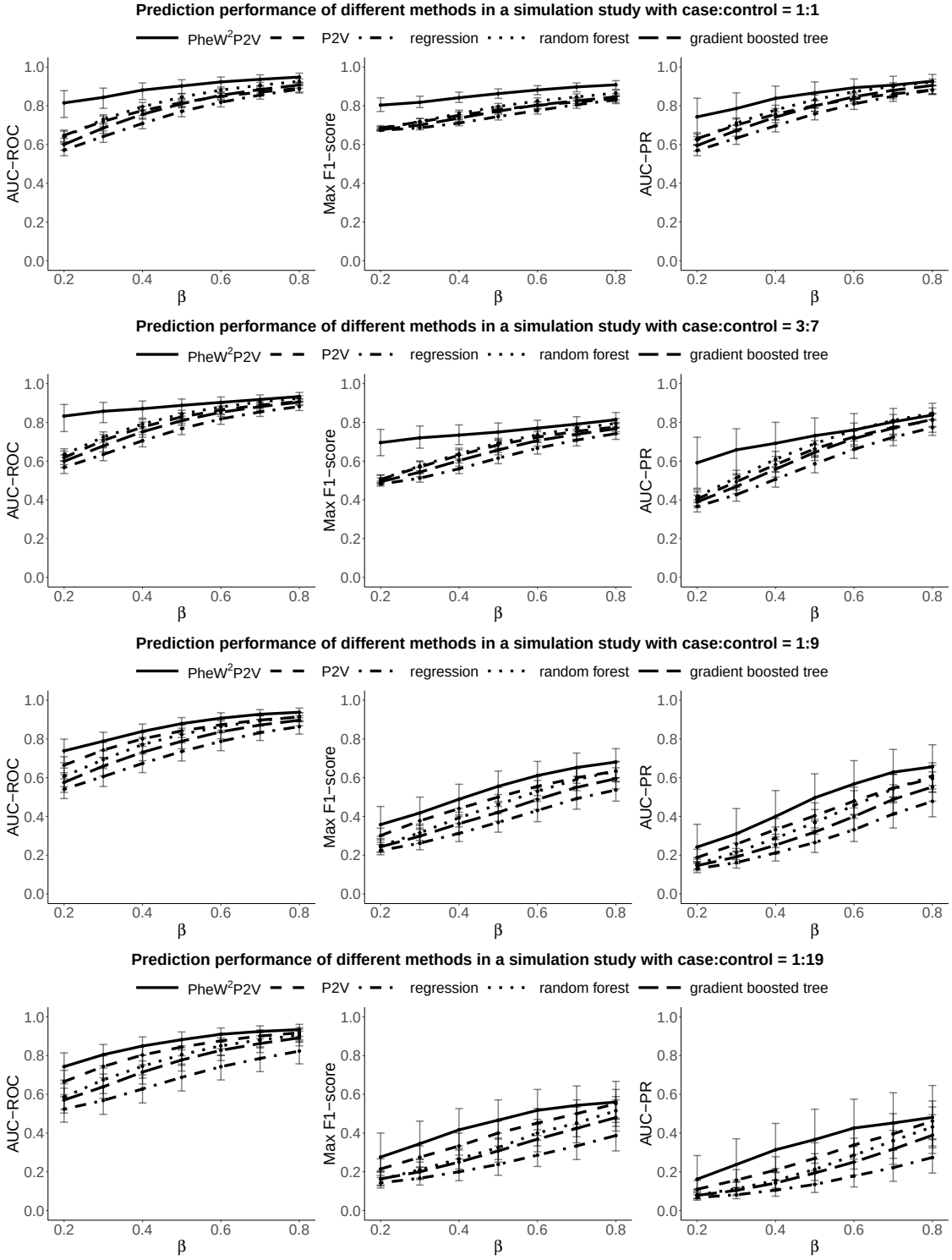

Figure B.3: Simulation results medians and 25th and 75th percentiles of AUC-ROC, max F_1-score, and AUC-PR of the proposed PheW^{2}P2V, the LASSO regression, the random forest classifier, the gradient boosted tree classifier, and the unweighted version P2V, with different case-control ratios of 1:1, 3:7, 1:9 and 1:19.

Appendix C: Appendix to multi-view graph convolutional clustering with applications to cancer subtyping with multi-omics data

C.1 TCGA STAD cancer data

Similar to the LIHC showcase, we used the R package TCGAbiolinks to obtain the multi-omics data for stomach adenocarcinoma (STAD) patient. There were 272 patients having gene expression data, 395 patients having DNA methylation data, and 431 patients having mutations data, resulting in 226 patients having all three types of omics data. We then removed 9 patients who did not have follow-up days information for prognosis, leading to 217 patients as our STAD patient cohort. Other quality control steps are the same as the LIHC showcase.

C.2 STAD subtypes identified by MultiGCC

Using the integrated multi-omics data, the proposed MultiGCC identified STAD subtypes, with a survival p-value 0.014 (main text Table 4.2), while subtypes identified by other comparison methods had non-significant associations with patient survival. The TCGA group (Network et al., 2014) identified four subtypes using 294 STAD patients using a consensus clustering method (Wilkerson & Hayes, 2010). But these four subtypes were not significantly associated with patient survival (p=0.89).

We displayed the Kaplan-Meier survival curves of the three STAD subtypes identified by Multi-GCC in Figure C.1A. There is a clear difference in patient survival across the three subtypes. For example, the subtype 1 with 52 patients has the worst survival and a median survival time 874 days. The subtype 3 with 40 patients has the best survival with a median survival time longer than 1223 days.

Figure C.1B and Figure C.1C displays the heatmap of selected gene expression and DNA

methylation profiles, respectively, where we selected top 500 genes/CpGs with smallest p-values from differential analysis using Kruskal-Wallis test comparing the three STAD subtypes identified by MultiGCC. We clearly observed different patterns across the identified subtypes. For example, subtype 2 has the highest gene expression levels at many Cancer Gene Census (CGC) and also STAD related genes such as *PTPRC* (Z. Liu et al., 2021), *CCR5* (Ryu et al., 2018), *CCR7* (Ryu et al., 2018), and *FLI1* (Del Portillo et al., 2019). Subtype 1 and subtype 3 have a higher DNA methylation level than subtype 2 at many CpGs that located in STAD related genes, such as cg02604211 on the gene *LGR4* (de Souza et al., 2019) and cg11344533 on the gene *SIK2* (Dai et al., 2021). In addition, subtype 3 has the highest DNA methylation level and subtype 2 ranks second at many CpGs that located in CGC and also STAD related genes, such as cg04009932 on the gene *TP53* (Fenoglio-Preiser et al., 2003). Figure C.1D displays the mutation landscape of the top 30 genes ranked by mutation frequencies. We also observed different mutation patterns across the three STAD subtypes. For example, 77% patients in the subtype 1 had *TP53* mutations, but only approximately 35% patients in subtype 2 and subtype 3 had *TP53* mutations. Mutations of another CGC and STAD related gene *ARID1A* (Inada et al., 2015) occurred in 43% of patients in subtype 3, but it occurred in only 15% of patients in subtype 1. Mutations in *LRP1B* occurred in 40% of patients in subtype 3, but only occurred in 22% in subtype 2, which is also a CGC and STAD related gene (S. Hu et al., 2021).

C.3 Hub gene analysis in STAD patients

Similar to the LIHC showcase, we investigated gene activities of differentially expressed genes across the three STAD subtypes identified by MultiGCC. Using the Kruskal-Wallis test for each type of omics data separately and with a Bonferroni corrected threshold, we identified 2,535 differentially expressed genes, 103,809 differentially methylated CpG sites, and 2 differentially mutated genes. We further selected top 200 differentially expressed genes, top 200 differentially methylated CpG sites which were mapped to 170 genes, and 2 differentially mutated genes. This leads to 368 unique genes. The protein-protein interaction (PPI) network and connection relationship of

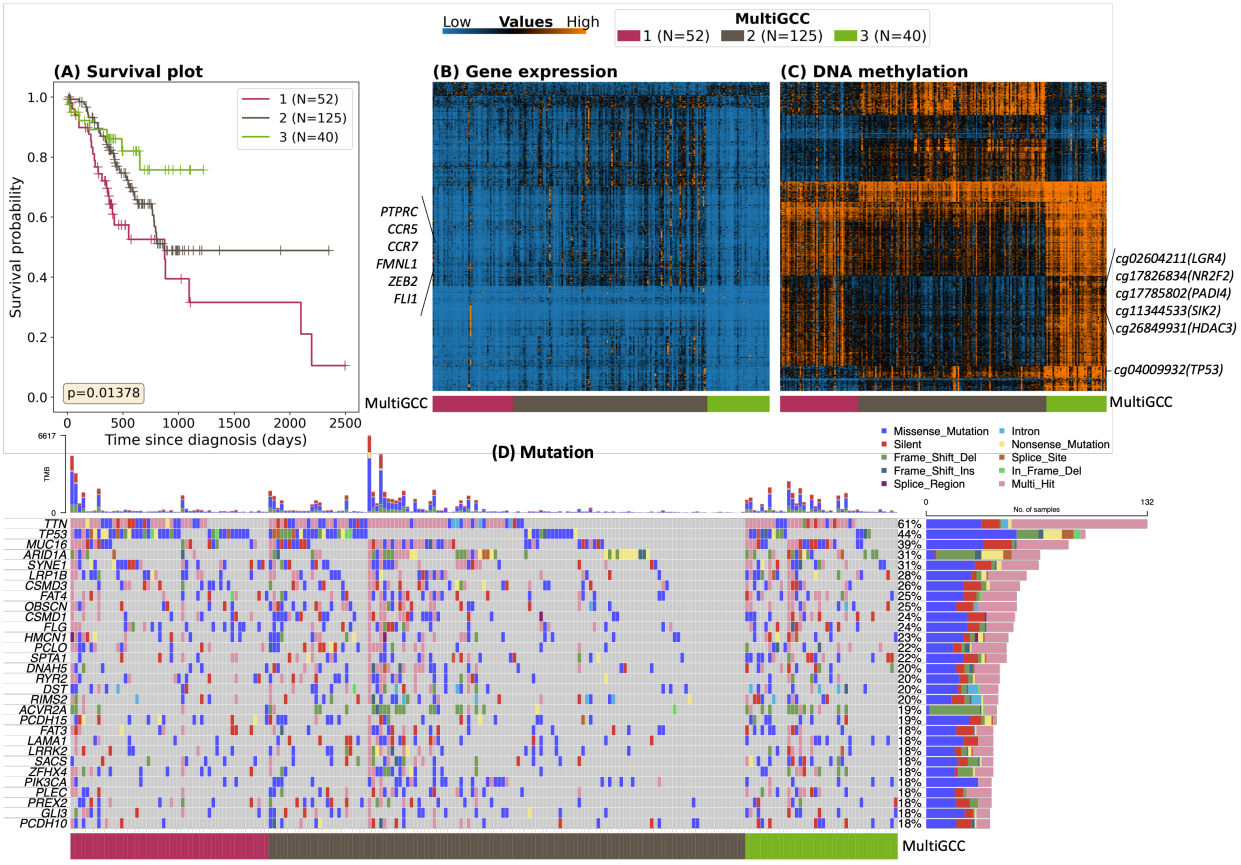

Figure C.1: Subtyping analysis of the three STAD subtypes identified by MultiGCC. (A) Kaplan-Meier survival curves and log-rank test p-value of the three STAD subtypes. (B) Heatmap of top 500 gene expressions that are differentially expressed across the three STAD subtypes by significance from the Kruskal-Wallis test. (C) Heatmap of top 500 DNA methylation CpG sites that are differentially methylated across the three STAD subtypes by significance from the Kruskal-Wallis test. (D) The middle chart displays the heatmap of mutation profiles of the top 30 genes ranked by mutation frequencies. The top chart displays absolute number of mutation load for each sample. The right panel displays the mutation frequencies.

the identified genes was retrieved from STRING database with a medium interaction score (interaction score ≥ 0.4). 326 genes out of the 368 genes were mapped to the PPI network with 922 edges.

Table C.1 lists the top five genes ranked by degree, stress, and betweenness centrality, in the PPI network of 326 genes. These top genes are known to be related to STAD and are also differentially expressed across the three LIHC subtypes identified by MultiGCC. For example, *PTPRC* had the highest value in degree, stress, and betweenness centrality. *PTPRC* gene is differentially expressed across the four subtypes with a Bonferroni corrected p-value<0.0001, with subtype 2 having the highest expression level (Figure C.1B). Recently, *PTPRC* has been found to be overexpressed in gastric cancer and is associated with the development of gastric cancer (Z. Liu et al., 2021). Gene *TP53* ranked second in degree, stress, and betweenness centrality. DNA methylation levels of CpG site cg04009932 in gene *TP53* was significantly methylated across the three subtypes with a Bonferroni corrected p-value<0.0001, where subtype 3 has the highest methylation levels (Figure C.1C). *TP53* is a well-known cancer-related mutation and is one of the most popular genes in cancer research (Guimaraes & Hainaut, 2002; Petitjean et al., 2007). Studies have found that *TP53* alterations occur early in the development of gastric carcinoma and could be a useful biomarker for cancer immunotherapy (Fenoglio-Preiser et al., 2003; L. Li et al., 2020). The gene *CCR5* ranked third in stress and ranked fourth in degress. Gene *CCR5* is also differentially expressed across the three subtypes with a Bonferroni corrected p-value<0.0001, where subtype 2 has the highest expression levels (Figure C.1B). *CCR5* and its ligand have been found to be differentially expressed in gastric cancer (Ryu et al., 2018).

Table C.1: Top five genes ranked by degree, stress, and betweenness centrality, in the PPI network of differentially expressed genes in STAD patients.

Gene	Degree	Gene	Stress	Gene	Betweenness centrality
PTPRC	61	PTPRC	69586	PTPRC	0.1694
TP53	41	TP53	59580	TP53	0.1684
IKZF1	38	CCR5	34810	ESR1	0.0789
CCR5	36	IKZF1	27610	IKZF1	0.0685
IL10RA	32	GNG2	24344	LRRK2	0.0675

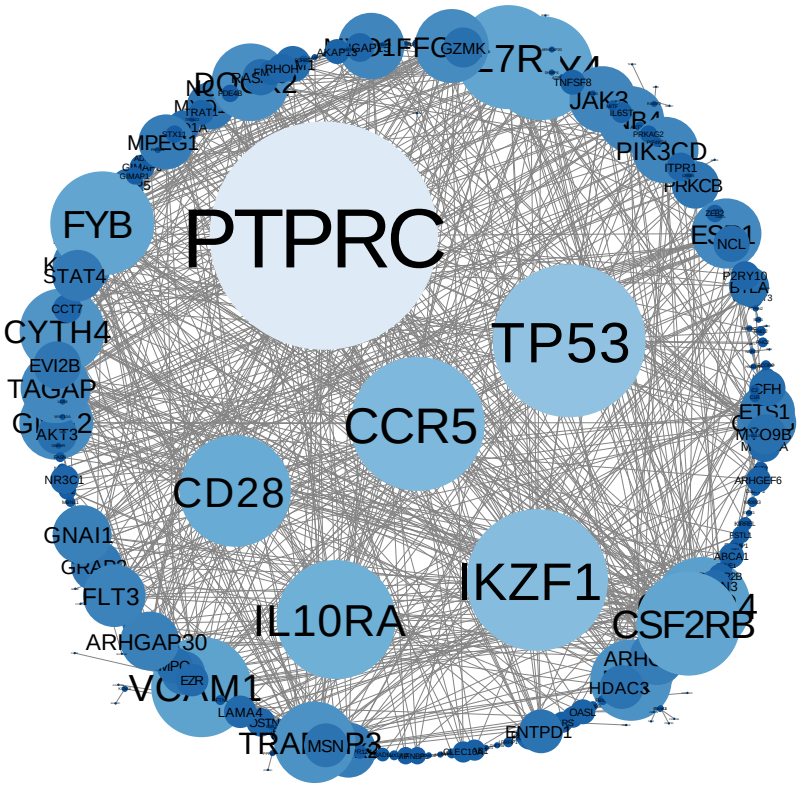

Figure C.2: Hub gene analysis of the 326 differentially expressed genes across the three STAD subtypes that were mapped to the PPI network. The color and size of each gene node represent the degree of each gene.