Use of the Electronic Health Records to facilitate phenotyping, comorbidity analysis, and genomics

Su Xian

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Peter Tarczy-Hornoch, Chair

David R. Crosslin

Sean D. Mooney

Gail P. Jarvik

Lea Starita

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

Use of the Electronic Health Records to facilitate phenotyping,

comorbidity analysis, and genomics

Su Xian

Chair of the Supervisory Committee:

Peter Tarczy-Hornoch

Department of Biomedical Informatics and Medical Education

Since the wide adoption of electronic health records (EHR) in 2010, many topics regarding the secondary use of the EHR received attention. The secondary use of EHR usually indicates repurposing the EHR data for research use, including information extraction, phenotyping, disease surveillance and forecasting, and policy making. Within this context, we ask how to use the EHR data to study the disease of interest, especially identifying new knowledge. In this work, we explored the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. In aim 1, we present an unsupervised approach for embedding high-dimensional EHR data at the patient level to help characterize patients and identify new disease patterns. Inspired by the modern language model architecture - transformers, with the attention mechanism - we use patient diagnosis and procedure codes as vocabularies and treat each patient as a sentence to perform the patient embedding. Using 34,851 medical codes for 1,046,649 longitudinal patient events, we performed embedding for 102,739 patients in the electronic MEdical Records and GEnomics (eMERGE) Network. In aim 2, we illustrated several downstream task applications of the patient embedding, especially providing insights into comorbidity patterns and the progressional trajectory of

individual patients within certain diseases of interest. We demonstrated excellent performance in the prediction of future disease events (median AUROC = 0.87, one year within the future), and bulk-phenotyping (median AUROC = 0.84). More importantly, we illustrated the use of patient vectors to reveal heterogeneity comorbidity patterns (disease subtypes) within a defined phenotype and captured their disease trajectory longitudinally. Our model is externally validated using the EHR dataset from the University of Washington, showing robustness and stable performance. These results paved the way for using representation learning in the EHR to characterize patients with certain diseases of interest and associated clinical outcomes that can promote disease forecasting performances and facilitate personalized medicine. In Aim 3, we utilized an EHR-derived and validated rule-based phenotyping algorithm to establish the cohort for identifying genetic risk factors for depression. We illustrated the application of genomic study using this EHR-derived algorithm to facilitate the study of disease etiology using genetics. We took a complex psychiatric disease -- depression, a leading cause of disability -- as an example, to study the genetic predisposition using data from the EHR. Large-scale genomic studies have identified common variants associated with depression. However, the complexity of the depression phenotype caused its suffering from inconsistent cohort definition and limited sample sizes. There is a need for a validated, automated EHR phenotyping algorithm that can accurately identify depression in the clinic. Here, we implemented a validated EHR phenotyping algorithm to construct a depression cohort (11,532 cases and 39,631 controls, total n = 51,163) and conducted a genome-wide association study (GWAS) using this cohort. Our study reproduced previously identified genetic associations (*PHF5A*, *KCNG2*) with depression susceptibility. We also identified novel SNPs falling into the *HLA* region and the *IGVH* region, indicating an association between the immune function and depression phenotype. In addition, we also demonstrated the robustness of our phenotyping algorithm through genetic correlation analysis, using a large meta-analysis of major depressive disorder as a standard. Together, this work served as a non-exhaustive but powerful demonstration of the use of the EHR data both in a supervised and unsupervised manner, to facilitate many downstream clinical applications, including phenotyping, comorbidity analysis, and genomics.

Table of Contents

Table of Contents **5**
List of Figures **7**
List of Tables **8**
Acknowledgement **9**
Dedication **10**
Chapter 1: Introduction **10**
1.1 Problem 11
1.2 Structure 13
1.3 Dissertation Aims 14
1.4 Contributions 16
Chapter 2: Background **19**
2.1 EHR Phenotyping 19
2.2 Representation learning 20
2.3 Language model 24
2.4 Genetics, precision medicine, and the potential of EHR facilitated genome-wide association study (GWAS) 28
2.5 EHR-derived rule-based phenotyping algorithm 29
Chapter 3: Language Models for Patient Embedding **31**
3.1 Overview 31
3.2 Related Work 33
3.3 Methods 35
3.4 Results 40
3.5 Conclusion 45
Chapter 4: High-throughput Phenotyping **48**
4.1 Overview 48
4.2 Related Work 49
4.3 Methods 50
4.4 Results 51
4.5 Conclusion 54
Chapter 5: Disease Onset Prediction **56**
5.1 Overview 56
5.2 Related Work 57
5.3 Methods 58
5.4 Results 59
5.5 Conclusion 63

Chapter 6: Heterogeneity Analysis in Disease Progression 65
6.1 Overview 65
6.2 Related Work 66
6.3 Methods 67
6.4 Results 71
6.5 Conclusion 86
Chapter 7: EHR derived Depression phenotyping algorithm and GWAS 89
7.1 Overview 89
7.2 Related Work 92
7.3 Methods 93
7.4 Results 95
7.5 Conclusion 105
Chapter 8: Conclusion 109
8.1 Summary 109
8.1 Future work 111
References 114

List of Figures

Figure 3.1 Illustration of the model architecture 38
Figure 3.2 Two-dimensional TSNE representation of patient embedding. Left panel is colored by patient age (the darker the older) and the right panel is colored by patient gender (pink to female, blue to male) 41
Figure 3.3 Performance (precision and recall) differences in patient events with different numbers of codes. 42
Figure 4.1 Step-by-step illustration of high-throughput phenotyping. 50
Figure 4.2 Performances on high-throughput phenotyping. Top right showing the boxplot of AUROC and bottom showing the relationship between sample size (case/control ratio) and AUPRC. On the right each boxplot represents AUROC distribution categorized by disease class according to phecodes. 52
Figure 4.3 External evaluation of the performances on disease onset prediction. The left panel shows the boxplot of AUROC categorized by disease class according to phecodes. The left panel shows the relationship between sample size (case/control ratio) and AUPRC. 53
Figure 5.1 Step-by-step illustration of disease onset prediction. 59
Figure 5.2 Performances on disease onset prediction. Top right showing the boxplot of AUROC and bottom showing the relationship between sample size (case/control ratio) and AUPRC. On the right each boxplot represents AUROC distribution categorized by disease class according to PheCodes. 61
Figure 5.3 External evaluation of the performances on disease onset prediction. The left panel shows the boxplot of AUROC categorized by disease class according to phecodes. The left panel shows the relationship between sample size (case/control ratio) and AUPRC. 62
Figure 6.1 Bayesian Information Criteria (BIC) curve for model selection. (a, b) for eMERGE cohort and (c, d) for the UW cohort 72
Figure 6.2 Clustering analysis identified subgroups with distinct comorbidity patterns in Systemic lupus erythematosus patients (n = 1806) from the eMERGE cohort. 73
Figure 6.3 Clustering analysis identified subgroups with distinct comorbidity patterns in colorectal cancer patients (n = 2837) from the eMERGE cohort. 75
Figure 6.4 Clustering analysis identified subgroups with distinct comorbidity patterns in SLE patients (n = 2546) from the UW validation cohort. 78
Figure 6.5 Clustering analysis identified subgroups with distinct comorbidity patterns in CRC patients (n = 3673) from the UW validation cohort. 80
Figure 6.6. Longitudinal analysis revealed progressional differences within each cluster group in CRC. . 84
Figure 7.1 Manhattan plot of the combined ancestry GWAS analysis. 96
Figure 7.2 Manhattan plot of the European ancestry GWAS analysis. 98
Figure 7.3 LD plot for the index SNP rs202207567 in European ancestry 100
Figure 7.4 eQTLs for the four leading SNPs identified from GTEx 101

List of Tables

Table 3.1 Performance of reconstructing original diagnosis and procedure codes. 41
Table 3.2 Evaluation of the is_same_patient and the is_next_event performance for the sentence-embedding model..... 44
Table 3.3 External validation of the is_same_patient and the is_next_event performance for the sentence-embedding model using the UW cohort 45
Table 3.4 Evaluation of the is_same_patient and the is_next_event performance for the sentence-embedding model using the UW cohort 45
Table 7.1 Re-identified genes that are associated with depression from GWAS catalog 96
Table 7.2. Genome-Wide significant SNPs summary statistics on MHC-II region 98
Table 7.3 Top hits of HLA association analysis summary statistics 103
Table 7.4. LDSC regression for genetic correlation analysis 105

Acknowledgement

I would first like to thank my mentors, David Crosslin and Sean Mooney, who are great researchers and inspiring mentors to work with. I benefited enormously from their research experience and supportive mentorship.

I also want to thank my committee members, Gail Jarvik, Lea Starita, and Peter Tarczy-Hornoch. Each of them made a great impact on my work and my PhD journey, providing valuable feedback and instructions. Specifically, I want to thank Peter Tarczy-Hornoch, who kindly took over as my committee chair for the last 8 months of my dissertation and put a significant amount of effort into making this all happen smoothly when my original chair (Sean Mooney) left UW to take a position at NIH.

I want to thank the people who helped me throughout my PhD journey, including Nic Dobbins, Patrick Davis, and Martha Horike-Pyne. They provided substantial help and inspiration to my work.

I want to mention my incredible appreciation to Dimitrios Morikis, Hannah Carter, and Maurizio Zanetti, who made my academic journey possible.

I also want to thank my parents for their love and support. None of them are in good wealth or authority; instead, they share a wonderful spirit of freedom and curiosity.

Dedication

This work witnessed the whole COVID-19 isolation. Thus, to all the isolated, as you venture through the journey of being, one day, people who share the same enthusiasm will appear and eventually connect. As said, *not everything was lost in the flow of time*.

“It's no different from building stations. If something is important enough, a little mistake isn't going to ruin it all, or make it vanish. It might not be perfect, but the first step is actually building the station. Right? Otherwise trains won't stop there. And you can't meet the person who means so much to you. If you find some defect, you can adjust it later, as needed. First things first. Build the station. A special station just for her. The kind of station where trains want to stop, even if they have no reason to do so. Imagine that kind of station, and give it actual color and shape. Write your name on the foundation with a nail, and breathe life into it.”

— Haruki Murakami,

Colorless Tsukuru Tazaki and His Years of Pilgrimage

Chapter 1: Introduction

1.1 Problem

Electronic health records (EHR) have been widely adopted in the United States [1]. The "meaningful use" of EHR, released by the Department of Health and Human Services, aims to improve the quality and efficiency of care [2]. To also facilitate clinical research, the potential of the secondary use of EHR has been explored and drawn large attention in recent decades. One of the major topics is EHR-based digital phenotyping [3–5]. The electronic MEdical Records and GEnomics (eMERGE) consortium has launched Phenotype KnowledgeBase (PheKB) as a digital phenotyping knowledge base that engages multiple sites of large hospitals and universities to share their phenotyping algorithms developed using the EHR data [6,7]. Though most of the algorithms stored in PheKB are rule-based and validated by domain expertise, there are also efforts towards developing machine learning algorithms for EHR-based phenotyping. Various machine learning and deep learning methods are applied to build EHR-based phenotyping algorithms, including Support Vector Machines (SVM), random forests, logistic regressions, and neural network architectures [8–10]. These efforts are marching towards a better characterization of diseases and aim to build a better healthcare system. Currently, a crucial and challenging question is how to leverage the EHR data to help identify and characterize disease patterns. With the ideal solution, we can promote disease monitoring and clinical predictive tasks and ultimately build a better healthcare system.

Representation learning is a powerful tool that can characterize existing and uncover novel disease patterns to facilitate the study of disease etiology, prevention, forecasting, and even heterogeneity analysis [11–14]. In the current era, the secondary use of EHR research heavily relies on existing domain knowledge from expertise [15–18]. Researchers in modern times are trying to find new patterns across the EHR data for different diseases [19]. There are existing findings using the EHR to unfold heterogeneity within a defined

phenotype, revealing differences in comorbidities potentially linked with genetics, lifestyle, and environmental factors [20–22]. Representation learning has emerged as one crucial tool dealing with high dimensional EHR data to facilitate pattern recognition, heterogeneity analysis, and downstream prediction tasks. In the EHR, one patient can have multiple visits in a year or across different years, generating abundant diagnosis and procedure codes accompanied by numerous lab values and observations, making it difficult to summarize. One typical representation of patients utilizing the nature of the EHR data structure is through binary vector representation, indicating if patients have specific diagnoses, procedures, or labs. Under these circumstances, matrix decomposition becomes a convenient tool to encode patients into relatively lower numerical spaces [23]. Matrix decomposition, such as principal components analysis (PCA) and non-negative matrix factorization (NMF), is used to compress a large matrix of patients into a relatively smaller one, while preserving the relationship of the selected features, such as diagnosis, procedures, and labs. The output matrix is usually termed patient embedding. Besides traditional matrix decomposition, a specific deep-learning model adopted from the autoencoder has been implemented to perform the embedding task [24]. Apart from the purely data-driven method, some methods integrate medical entities to perform predictive tasks [25]. In short, without the obligation of cohort building or domain expertise to iteratively process the phenotyping tasks, various unsupervised methods are applied to identify unknown patterns of patients and diseases using the EHR data.

Genetics is another area that can be greatly benefited from the adoption of the EHR [26–28]. The integration of genetics and EHR represents a powerful intersection of biology and technology, unlocking new frontiers in personalized medicine. EHRs provide a comprehensive digital repository of patient information, including demographics, medical history, laboratory results, and treatments. When combined with genetic data, this synergy allows researchers and clinicians to uncover the genetic underpinnings of certain refined phenotypes or diseases. The incredible large-scale and diverse data within the EHR enables population-wide genomic studies, enabling thorough explorations into rare and complex diseases [27]. Further, by

harnessing the combined potential of genetics and EHRs, researchers can accelerate discoveries that lead to improved healthcare outcomes and facilitate translational bioinformatics under the clinics [29,30].

1.2 Structure

Chapter 1 serves as a broad introduction to the work, expanding from the research questions to each specific aim with a general summary of the contribution of this work. Chapter 2 provides a solid but non-exhaustive background for the readers to understand the work better.

Then, from chapters 3 to 6, we focused on a data-driven, unsupervised representation learning of the EHR, to build modern tools to facilitate the study of diseases using EHR data. In chapter 3, we explored the potential of using medical diagnosis and procedures to construct patient vectors. In the first few sessions of this work, we represented each patient as a sentence, using the medical diagnosis and procedures codes as vocabularies to compose patient vectors. Moreover, we generated patient vectors in a longitudinal format, allowing investigation of a patient at a specific time point. In chapters 4 and 5, we showcase the language-model-based EHR patient data embedding and utilize its power in bulk phenotyping, and future disease prediction. chapter 6 is a detailed exploration of novel comorbidity studies. In this chapter, we demonstrated that cluster groups within a certain defined phenotype can have different disease progression trajectories longitudinally, which is crucial for understanding the causality and etiology of disease subtypes.

Then, from Chapter 7, we focused on illustrating a rule-based phenotyping algorithm for depression, a complicated psychiatric disease. We demo an EHR-based depression phenotyping algorithm to perform a genome-wide association study. This analysis revealed a practical use of manually developed phenotyping algorithms and served as a solid genetic validation of the algorithm. This chapter seems to be independent of the previous work. Nevertheless, this chapter is a refined analysis of the standard secondary use of EHR

information to facilitate genomic studies. Though previous chapters are pioneer works and data-driven, this session presents the use of expertise-based methods to unravel the information from the EHR.

1.3 Dissertation Aims

In this dissertation, we explore the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. To identify novel patterns of diseases and investigate diseases' etiology, we approach this question with both an unsupervised approach and a supervised approach. We developed and assessed a language-model-based unsupervised learning approach for disease pattern identification in Aim 1 and Aim 2. In Aim 3, we utilized a supervised (rule-based) depression phenotyping algorithm to construct a cohort for identifying genetic risk factors, as connecting genetics with molecular functions can provide insights into disease etiology.

1.3.1 Aim 1. To develop patient representation learning in EHR data using an unsupervised machine learning approach.

One of the biggest challenges for EHR data analysis is high dimensionality and data sparsity, as the EHR is not designed for research purposes. The current use of EHR data requires domain expertise in using their knowledge to form specific hypotheses and perform certain types of analyses. To overcome the high-dimensional curse, we first aim to develop an embedding algorithm representing patients in numerical space to perform downstream calculations. The majority of the current efforts and methods focus on matrix decomposition technologies for patient embedding utilizing EHR data (Becker et al. 2022). We propose a language-model-based embedding method, which utilizes the attention mechanism targeting patient embedding [31]. We will use the longitudinal EHR data, taking diagnosis codes, procedure codes, and medications as the basic building blocks for patients to perform the embedding.

1.3.2 Aim 2. To identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors.

Comorbidity is crucial for personalized treatment and progression [32]. We use the embeddings to perform clustering analysis (Gaussian mixture model) and identify stable clusters within a single phenotype that can reflect their comorbidity heterogeneity. In this work, we use colorectal cancer (CRC) and systemic lupus erythematosus (SLE) as two instances, demonstrating the use of embeddings to identify heterogeneity of comorbidity patterns within a single phenotype. Furthermore, we show that using these longitudinal vectors of patients, we can identify different trajectories of disease progression and analyze the specific progression patterns and related health outcomes. We show that our patient embeddings can better characterize disease heterogeneity and facilitate personalized treatment.

1.3.3 Aim 3. To perform a genome-wide association study (GWAS) for depression phenotype using rule-based phenotyping algorithm derived from the EHR.

Depression is a complex phenotype with multifaceted constructs, including a wide range of symptoms, behaviors, and biological markers. We adopted the phenotyping algorithm for depression developed by the eMERGE consortium [33], performing Genome-Wide Association Studies (GWAS) to identify genetic risks associated with depression. In this work, we illustrate using an EHR-derived phenotyping algorithm to facilitate novel genomic discovery. Meanwhile, we use the genetic correlation and GWAS findings to compare and validate the rule-based depression algorithm. Further, with an interest in depression and immune phenotypes, we will focus on the Human Leukocyte Antigen (HLA) region, examining the immunological and inflammatory response aspect and providing new insights about the disease risk and potential etiologies.

1.4 Contributions

We explored the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. In the first two aims, our novel model architecture demonstrated an effective embedding of EHR data at the patient level and showcased a few downstream applications. To our knowledge, no multi-purpose flexible model has been developed, as most studies use complex model architectures focusing on a single downstream application. In the third aim, we took the traditional rule-based phenotyping approach, performed a genome-wide association study, and illustrated the genetic validation of the algorithm. Together, this work served as a combination of an innovative approach (unsupervised approach) to data-driven EHR analysis, and a traditional guide (supervised approach) for genetic studies and rule-based phenotyping algorithm validations under the clinic.

Many studies have developed novel models to fit the EHR data. Our approach utilizes modern neural network architectures, first addressing the security concerns by embedding patients into numerical space, which not only removes identifiable personal information but also overcomes the high-dimensional challenge of summarizing patient information. Consequently, these numerical vectors first serve as features for patients and then provide values in de-identifications and systematic visualizations. Aim 1 adopted the current state-of-the-art language model to perform patient-level embeddings [31]. In the evaluation phase, we ensured the embedded patient vector can reconstruct the original patient information (e.g., diagnosis code, procedure codes, medications, etc.). This step ensures the model quality of the embedding.

Because the embedding utilizes an unsupervised learning method, the learned information about patients is not limited to existing understandings of expertise. In the development of medicine, many experiments serve to understand human diseases and study their progression and treatment. However, there are still many unknown disease patterns and unsolved puzzles. Unsupervised learning is a powerful tool that can

capture delicate patterns which are not penetrable by humans. That is to say, unsupervised machine-learning models are much better at identifying patterns than humans [34]. Aim 2 revealed the value of the embeddings by performing several downstream tasks. We will first demonstrate that the embeddings can achieve great performance on standard but clinically meaningful tasks such as disease onset prediction and patient phenotyping. Then, we will illustrate its power in revealing unrecognized patterns within diseases or phenotypes and provide new insight into personalized medicine.

Most importantly, this work exhibited a novel data-driven EHR approach to study disease patterns. To our knowledge, traditionally, researchers only explored a small fraction of the EHR data based on their needs or scope [15,35]. It is understandable, as EHR data analysis is usually expertise-driven, which means only people with domain knowledge are currently interested in exploring the research value of the EHR. Besides, accessing EHR requires a series of complicated administrative processes and intensive training to ensure data security, such as de-identification processes to protect patient privacy and data integrity. Thus, though EHR data might have great potential and research value, many barriers remain in front of the practical use of these data. Our work can provide another layer of safety to patients, as each is encoded into numerical spaces without revealing personal information.

The third aim addressed the other side of the coin, using traditional rule-based (supervised) phenotyping algorithms to facilitate large-scale genomic studies, leveraging the EHR data. Besides the popularity of machine learning patient phenotyping algorithms, domain expertise uses EHR data to design rule-based algorithms that often integrate powerful knowledge and the most updated information. There are many examples of employing expertise-designed EHR phenotyping algorithms to facilitate genomic analysis. Aim 3 focuses on a complex disease, depression, which has been studied primarily as a psychiatric disease [36]. Our goal is to identify potential genetic risks for depression patients using GWAS. Specifically, utilizing the phenotyping algorithm and combined genotype data, we will explore the hypothesis that originated ages ago, explaining the immune-dysfunction component of depression, which sheds a new path

toward aspect for us understanding depression for depression from a new perspective, providing values for both the study of etiology and the molecular pathways [37–39].

Together, this project served as an exploratory EHR data analysis, focusing on investigating novel disease patterns and disease etiology, leveraging modern deep learning techniques and traditional knowledge-based methods. We first applied the modern architecture of deep learning – language models to perform high-dimensional numerical embeddings of patients using the EHR data. Then, we illustrated the practical use of these embeddings for standardized tasks such as disease onset prediction and patient bulk phenotyping. Specifically, we identified unexplored disease patterns using clustering analysis and patient comorbidity heterogeneity analysis. Last, to gain insights into disease etiology, we utilized a domain expertise-designed depression phenotyping algorithm from the EHR to construct the cohort and identify genetic risk factors.

Chapter 2: Background

This chapter serves as a basic knowledge pool to help readers better understand the work we did to explore the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. The structure of this chapter can be divided into two parts. The first part starts from 2.1 EHR phenotyping and ends in 2.3 Language model (including sub-chapter 2.3.1 Transformer model and attention mechanism). The first part prepares the reader for the first two aims of this dissertation, which are described from chapter 3 to chapter 6. The second part starts from 2.4 genetics, precision medicine, and the potential of EHR-facilitated genome-wide association study (GWAS), to the very end of chapter 2. The second part adds all the preliminary knowledge and history of the third aim, described in chapter 7.

2.1 EHR Phenotyping

EHR is an abundant and rich place full of potential for biomedical research. One of the critical tasks within the application and study of the EHR is called digital phenotyping (or EHR phenotyping). Digital phenotyping utilizes information on patient data from the EHR to define or characterize patients into groups, cohorts, or a phenotype [3,40]. Before the wide adoption of machine learning technology, conventional digital phenotyping leverages knowledge from domain expertise that manually searches patterns in diagnoses, procedures, medications, clinical notes, sometimes radiology reports, etc. Conventional digital phenotyping algorithms often are in the form of pseudo-code, including descriptive text and flow charts illustrating the steps of algorithms. While machine learning algorithms are mostly composed of computer programs and specific statistical models with selected features from the EHR to compute the probability that patients have a phenotype or not. With the advent of more efficient machine learning algorithms and great computational power, bulk-learning, or high-throughput phenotyping has come to the attention. Bulk-learning, or high-throughput phenotyping are tasks that include using a great

amount of EHR data and features to perform patient characterizations that can output thousands of phenotypes. Though the conventional phenotyping approach requires a great effort and manual process, which usually goes through iterations of correction to produce the final pseudo-code for a phenotyping algorithm, the performance is comparable to and sometimes even outperforms modern machine learning approaches, making it still a popular choice in phenotyping tasks.

2.2 Representation learning

Representation learning is a method that can be categorized into unsupervised machine learning. The core concept of representation learning is to learn the representation of data [13]. For example, image compression algorithms that can compress a large image into a smaller size but retain the crucial features of the original image, is a form of representation learning. Besides, modern representation learning can also be applied to abstract objects, such as words, sentences, and paragraphs. Word representation learning, commonly known as word embedding, is a method that can represent each word into a high-dimensional numerical space [41,42]. Through these embeddings, words that have similar meanings or co-occurred oftenly usually form patterns, such as having closer Euclidean distances [42]. Some models went a step further and performed sentence embeddings that projected each sentence into a high-dimensional numerical space so that sentences with similar meanings have a closer Euclidean distance.

2.2.1 Neural network

Nowadays, due to both the advancement of computational power and the efficient design of backward propagation functions, the neural network has become one of the biggest achievements in the field of machine learning and artificial intelligence [43]. A neural network is a computational model inspired by biological neural networks in the human brain processing information. It consists of interconnected nodes, or "neurons," organized into layers. Neurons are the basic computing elements within the neural network.

Typically, a neuron receives one or multiple inputs from other neurons in the previous layer, processes them, and produces an output, which is directed to other neurons as inputs. The processing typically involves a weighted sum of the inputs followed by an activation function. Mathematically, the computation process within a single neuron is a simple linear function, illustrated below:

$$Wx + b = y$$

With *W* as an array indicating weight parameters, *x* an array representing inputs from previous layers, *b* as the bias term. As we know, any combination of linear functions can also be represented in a single linear function. Thus, a neural network without activation functions is not complex enough and can be just represented as a linear equation, losing the meaning of building a complex neural network.

While the activation function can have many forms, depending on the need and sometimes data type. Some common activation functions include sigmoid, tanh, ReLu (rectified linear unit). The activation function *a* is added to the output *y*, to ensure an non-linear transformation of the neuron, illustrated as below:

$$f = a(y)$$

A typical function of *a* can be ReLu, which simply takes the maximum value between the output *y* and 0, illustrated below:

$$f = max(0, y)$$

Neurons, as described above, form layers together, and each layer is connected to the neighboring layers, to transmit the computational signals. Usually, the first layer is called the input layer, and the last layer is the output layer. The in-between layers are usually hidden and not of main interest, thus called hidden layers. Hidden layers can range from a few to tens and hundreds, depending on the complexity of the neural network design.

To make the neural network model trainable, the weight parameter *w* within each neuron will be adjusted in each training step. During the training, all computations flow from left to right and ultimately will yield an output value from the output layer. Then, the discrepancy between the model output value and the true

value is measured, called loss. There are numerous loss functions to compute the loss, with each fit into different scenarios of model training and data type. Typically, for binary label prediction (1 versus 0), the model will adopt a binary entropy loss function. While the mean squared error function is commonly used for continuous value prediction (similar to regression tasks).

After calculating the loss of a model, backpropagation will be applied to tune the weight parameter *W* of each neuron in a backward fashion (right to left) [44,45]. The backpropagation algorithm is a crucial component of neural networks. It is no exaggeration to state that the significant success of modern neural networks is largely attributable to the elegance of the backpropagation algorithm. As previously illustrated, neurons are interconnected to produce outputs, necessitating the adjustment of each neuron's weight W to align computed results more closely with actual results. Backpropagation facilitates this feedback loop from the calculated output to each neuron's weight. Specifically, it calculates the differences between the expected and computed outputs (the loss) and then derives gradients with respect to each neuron's activation function. This process indicates the direction in which the weights should be adjusted to minimize the loss and bring the computed output closer to the expected output. You might notice that the derivative only tells us the direction we need to go (and a certain magnitude, yes) but with no certainty, the model will arrive at a global minimal point that perfectly minimizes the loss. Therefore, a learning rate η is usually introduced in the model to adjust the magnitude of weight change during each iteration of backpropagation. For example, and updated weight W_* is calculated by (where L indicate loss):

$$W_* = W + \eta \frac{dL}{dW}$$

Another common question neural networks face is that, since the parameter update (weights) are somewhat arbitrary, it is never guaranteed to achieve a real global minimum. In fact, there is no need for a neural network to achieve a global minimum, as it can almost efficiently reach extremely close local minimal loss, which is good enough. This property is also a great proof of the versatility and flexibility of neural networks.

2.2.2 Deep learning

Deep learning is a subset of machine learning that involves the use of neural networks with many layers, often referred to as deep neural networks. These networks are capable of learning from large amounts of data, identifying patterns, and making decisions with minimal human intervention.

Initially, deep learning consists of three basic neural network types, including dense neural network (DNN, sometimes also called feed-forward neural network), convolutional neural network (CNN), and recurrent neural network (RNN). Three of each have different applications and focuses. DNN is the most traditional neural network that is similar to what we've mentioned in the above session, which consists of densely connected neurons in each layer and can be used to perform regression or classification tasks. CNN is efficient in image processing, as it applies filters that can learn detailed structure and features of a picture, which has been widely implemented in computer vision. RNN is intrinsically different from DNN and CNN, with a complicated model architecture specifically developed for sequence data.

Within the most recent decade, some novel model variations came to our sight. Autoencoder, a DNN-based symmetric neural network, is developed to perform compression and embedding tasks with the power of learning non-linear features. Transformer, a novel attention-mechanism-based language model built upon feed-forward neural networks, surpassed RNN in computational efficiency due to its parallel input options. The transformer is demonstrated to be superior in several downstream tasks, specifically in pairwise translations.

2.2.3 Autoencoder model

An autoencoder is a type of artificial neural network used for unsupervised learning. Its primary purpose is to learn efficient codings of input data, typically for dimensionality reduction or feature learning [46,47]. Typically, an autoencoder consists of an encoder and a decoder. The encoder compresses the input data into

a lower-dimensional representation called the "latent space" or "bottleneck". The decoder takes the latent space and reconstructs the original input data. Autoencoders are powerful tools in unsupervised learning, capable of learning efficient representations of data (in the latent space) for various tasks. Their ability to compress and reconstruct data makes them versatile for applications in dimensionality reduction, feature learning, denoising, anomaly detection, and more [46].

2.3 Language model

A language model is a computational model designed to understand, generate, and manipulate human language. These models are crucial in a variety of natural language processing (NLP) tasks, including text generation, translation, summarization, and sentiment analysis. The primary goal of a language model is to predict the likelihood of a sequence of words and to generate coherent text.

Back in that time, language models were developed and mostly used to analyze discretized text data, such as the use of N-grams [48,49]. Gradually, more complicated statistical language models have been developed, such as latent Dirichlet allocation (LDA), used for topic analysis [50]. However, the above methods represent each word as a simplified token and usually ignore the sequential relationship of words in sentences or paragraphs.

Until recently, the word embedding technique advanced, and more complicated language models utilizing neural networks have been developed. Word embedding is a numerical representation of words that tries to encode information about words into the numerical space [41,42]. In fact, word embedding is no new concept and has been a study topic from distributional semantics since 1950 [51,52]. That is to say, though word embedding nowadays is usually performed by deep neural networks, old techniques exist for word embedding tasks and heavily rely on statistical language models. Popular word embeddings using neural networks including Glove and Word2Vec, which are trained based on large internet text and public

resources across all domains. The beauty of word embeddings can be illustrated in the figure below, where words with similar meanings are clustered tightly within the numerical space. Sometimes, some word embeddings can represent analogy in arithmetic ways, known as the famous "king - man = queen - woman" example [42]. There are also specific domain word embedding models, such as Med-BERT, which are optimal for downstream tasks within the medical domain [53].

Nowadays it is common to use neural networks with a tremendous number of parameters to train language models -- known as large language models, LLM --, with the help of efficient word embeddings as input. Famous LMM includes BERT, GPT, and so on. However, before we jump into the large language model era, we need to mention the sequence model, a specific type of neural network for dealing with sequential data, including Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models [54,55]. These neural networks are extremely complicated and very computationally intensive, due to the nature of feeding the input sequentially. However, these models have achieved state-of-the-art performance and were prominent until the advent of the transformer model [31].

Transformers initially introduced in "Attention is all you need", designed for pair-wise sentence translation, invented the attention mechanism, allowing a sequence of specific length as input which achieved a huge computational efficiency improvement compared to RNN-like models [55]. Moreover, the attention mechanism was then found to have the ability to capture crucial feature interactions among words. Thus, without any doubt, the transformer becomes the new king in the modern language model field.

Language models are nowadays popular in many fields and are also widely used in the clinical domains [56,57]. Language models can be used to extract information from clinical text, detect symptoms, or build predictive models for certain clinical outcomes [58]. Most importantly, Language models have achieved several notable challenging medical tasks and inspired novel model development within the medical field [53,59].

2.3.1 Transformer model and attention mechanism

The detailed transformer model architecture can be found in the original work "Attention Is All You Need" [31]. We will give a short introduction of the model architecture to prepare for the utilization of the attention mechanisms in the following sessions. In general, the model consists of an encoder part and a decoder part. As the transformer model is designed for pair-wise translation, the inputs are a pair of sentences in different languages, while the output is a probability matrix that denotes the probability of words in corresponding positions of the sequence. The encoder is composed of six identical layers, within each, two sub-layers. The two sub-layers include a self-attention layer and a feed-forward neural network layer.

The attention mechanism (i.e. the self-attention layer) utilizes matrix operations to compute the “attention score” that captures the relationship between input features. The mathematical formula of the attention function is:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{dk}})V$$

As we can see, there are three crucial players, matrix Q (query), K (keys), and V(values). These three vectors are derived from the inputs and are optimized during the training. The *dk* variable represents the dimension of K (keys), and we can see that the formula above is scaled by this parameter to avoid a large dot product causing vanishing gradient issue.

The above attention function is then termed "attention head", which are concatenated together to form multi-head attention. The mathematical representation is given below:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W$$

While each $head_i$ is:

$$\mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

The number of attention heads is denoted by the parameter *h.* In the original work, the author defined *h* = 8 to train the transformer.

After illustrating the above detail, please refer to the Figure 1 of the original manuscript to have a closer inspection of how the multi-head attention mechanism is used for both the encoder and decoder to perform the pair-wise translation work. Though the original work is designed to perform pair-wise sentence translation through the cooperation of the encoder and decoder function, the encoder and decoder architecture have evolved later to fit other tasks of interest.

BERT -- bidirectional encoder representation from the transformer, is a novel model that stacks the encoder of the transformer model to perform conceptualized word embedding [31,60]. Through this example, researchers gained a deeper understanding of the encoder architecture, realizing how the "latent space" produced by the encoder can maintain crucial information from input, serving as an embedding itself.

2.4 Genetics, precision medicine, and the potential of EHR facilitated genome-wide association study (GWAS)

GWAS is one of the most successful tools in modern genetic trait association studies. GWAS emerged about two decades ago and has identified many disease and gene associations. Nowadays, diseases are widely known to have an association with genetic factors, and some of them are caused by single gene mutations (monogenic), called Mendelian diseases. However, many complex traits/diseases are associated with multiple genetic factors (polygenic), and can also be influenced by environmental and behavioral

factors. Unlike Mendelian diseases, complex diseases are polygenic, which indicates that each gene only contributes to a small effect of disease onset.

GWAS study scans the whole genome to identify associations between single nucleotide polymorphisms (SNP) and traits of interest. As complex diseases have polygenic properties, the effect size of each SNP is too tiny and thus usually requires a large amount of data to reach genome-wide significance.

To reach populational level genetic studies that can further increase the power of GWAS analysis, EHR and large biobanks are indispensable. In 2006, the United Kingdom started the UK Biobank project in the long term, aiming to facilitate large population-level genetic predisposition diseases study. Likewise, approved by President Obama, the United States launched the All of Us Research Program, aiming to collect health data from at least a million people within the United States, to accelerate large population-level diseases and genetic association studies. All the above efforts are laying the foundations for a future of precision medicine that accounts for individual-level variabilities.

Precision medicine is an innovative approach to medical treatment that takes into account individual variability in genes, environment, and lifestyle for each person. Precision medicine aims to tailor medical care and interventions to the uniqueness of individuals, as opposed to the traditional "one-size-fits-all" approach. Genetics is one of the major focuses of precision medicine. Researchers have identified many causal variants (usually single nucleotide level variations) for human diseases, such as BRCA1 and BRCA2 for breast cancer [61]. Numerous efforts have been put into the clinic to provide early mammograms and on-time prevention of patients carrying such variants, as the carriers have higher risks of developing cancers than normal populations.

2.5 EHR-derived rule-based phenotyping algorithm

To better implement precision medicine in the clinic, large population genetic studies are necessary to unfold genetic variants that contribute to disease risks, especially for complex traits when multiple genes are associated with disease risks but have small effect sizes. Therefore, identifying a large population with a defined phenotype at high quality becomes a demanding task.

Traditionally, to conduct a GWAS analysis, researchers need to recruit participants and collect DNA from them. However, it is challenging for individual studies to recruit a significant number of participants to simulate a population-level genetic study.

Electronic health records (EHR), a system within the hospital that documents information from all encountered patients, raised an opportunity. If all hospitals started to collect DNA from their patients and build a large bio-repository for all this data, the sample size could easily reach a populational level. The University of Pennsylvania and some pioneers have launched this ambitious action of building a paired bio-repository for the EHR system [62,63]. An important question next is how we accurately identify patients with certain phenotypes from such a large hospital cohort.

This is where digital phenotyping tasks gained attention. Digital phenotyping, also known as computational phenotyping, utilizes computer programs to analyze digital information from the EHR in order to assess whether a participant meets the criteria for specific phenotypes. The eMERGE network started an initiative called PheKB, short for Phenotype KnowledgeBase, which stores the pseudo-code or documentation of computational phenotyping algorithms [6]. Though PheKB majorly documents rule-based phenotyping algorithms, novel machine-learning algorithms have also been implemented to perform phenotyping tasks. Machine learning algorithms are generally faster and require less iterative process of laborious manual evaluation, compared to traditional rule-based phenotyping algorithms [9,64]. However, there is no

evidence of any performance differences between these two methods [9]. As a result, both are widely adopted and applied in the research domain.

Chapter 3: Language Models for Patient Embedding

3.1 Overview

Broadly, this dissertation explores the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. To start, we focused on designing tools leveraging modern computational models to study disease patterns. To uncover novel disease patterns in the EHR, we proposed modern deep learning models, specifically language models, to approach this question. Our goal here, as stated in Aim 1(to develop patient representation learning in EHR data using an unsupervised machine learning approach), is to build a patient embedding model by adopting language models that represent patients in numerical space, alternatively, also called patient representation learning.

Patient representation learning, also can be called patient embedding, is a numerical representation of patient information using data from the EHR [11]. Patient embedding can learn and encode rich information about individuals, depending on the scope of features included within the embeddings. But still, what level of the information can be recovered and how accurate it is are heavily dependent on the quality of the patient embedding models. Though patient embedding, similar to word embedding or sentence embedding, is an abstract work that is not generally understandable at first glance, it has huge potential, especially powerful in the big data era. In this work, we will use patient representation learning and patient embedding interchangeably for convenience.

Patient embedding as an innovative way of unsupervised learning using EHR data, can capture nuances of patients and further facilitate the understanding of patient heterogeneity and fine-tune cohort building [14]. It is crucial to explore new features and heterogeneity to refine a phenotype and gain new insights to understand the comorbidity within, which can provide insight into personalized medicine. Moreover,

patient embedding can learn new patterns that are hard to capture simply through manual effort, as nowadays big data is intrinsically complex and high-dimensional. For example, a research group has explored the heterogeneity of a defined phenotype, using traditional topic-modeling techniques, latent Dirichlet allocation (LDA), and Poisson Dirichlet model (PDM) to perform patient clustering and identify survival discrepancies within clusters [14].

Though the concept of patient embedding is utterly fresh, modern neural network models emerged and demonstrated the power of patient embeddings. For instance, besides the traditional embedding method, autoencoder neural network architecture has been employed to perform a patient embedding task on almost all EHR features, from diagnosis, procedures, medications, and labs to clinical notes and basic demographics of patients [11]. Moreover, more advanced models, such as convolutional neural networks (CNN) and customized multilevel prediction from a sequence of events have been developed to perform medical embeddings, showing a growing interest and attention within the field [12,65]. The next step is to gauge the downstream applications within the clinical context.

At the present time, several neural network architectures have been developed and proven to show progress in various kinds of health-related tasks, both supervised and unsupervised [66,67]. Specifically, transformer attention mechanisms have surpassed the traditional sequence models, such as recurrent neural networks (RNN) and some of their variations (GRU, LSTM), both in efficiency and performance in downstream tasks within the natural language processing (NLP) field [60,68]. Several novel model architectures are built upon attention mechanisms, including GTP, BERT, and variations of BERT (such as Med-BERT, Bio-BERT) [53,59]. Though language models such as BERT are initially built for general use purposes, variations like Med-BERT demonstrated language models within a specific domain (for here the medical domain). Moreover, though language models are initially designed for processing text, they can be adopted for various purposes that only require the input data to be a sequence. Therefore, we adopted the transformer model with patient visits as sequences to perform the embedding.

In this session, we will describe the development of our patient embedding model and the initial results achieved through the patient embedding model. We will discuss the model performance and analyze the results to provide insight into future direction and downstream applications of patient embedding models. This session serves as the core of the first six chapters, as the model itself is responsible for producing meaningful patient embeddings.

3.2 Related Work

Till now, there are not many patient embedding works within the community. However, several models and concepts that arose are similar to patient embeddings. Since 2016, after the work of Miotto et al. [11], many others have used traditional clustering or topic-modeling techniques to perform patient heterogeneity analysis using EHR data [14,69]. Recently, advanced and complicated models were developed according to customized needs and served different clinical purposes. Landi et al developed a convolutional neural network framework, based on the previous work in 2016, to characterize patients of subgroups, focusing on 8 complex traits [12]. Choi et al. utilizing sequences of visits and levels of symptoms to treatment, developed a multilevel medical embedding using the EHR [65]. Ramsy et al. developed Med-BERT, which built upon the BERT model but specifically focused on the clinical domain to facilitate disease prediction [53]. As mentioned, almost all the work above are specific forms of patient representation learning or patient embedding that are tuned according to their needs. In this work, we adopt some basic concepts and follow the most traditional patient embedding models mostly similar to the work from Miotto et al. to develop our patient embedding models. Besides these traditional concepts, our model will be focusing on a new aspect that no other has tried before – focusing on the progression of disease and comorbidities to perform a longitudinal patient embedding.

Numerous efforts are exploring different methods of EHR embeddings, including traditional matrix factorization, autoencoder-based neural networks, and graphical representations (see Preliminary study

session). However, most existing tools have two issues: 1. Lack of consideration of longitudinal properties of the EHR. Most of the methods mentioned above (excluding some graphical representation methods) do not reflect the property that EHR data is naturally longitudinal, hence ignoring the fact that events in the EHR happen in a sequence of specified orders. Without encoding the longitudinal information, the embedding would lose the details about the timing of events. For example, a 30-year-old patient diagnosed with diabetes who developed COPD at age 60 is dissimilar to a 58-year-old diabetic patient who developed COPD at age 60. 2. Most contemporary famous embedding methods fail to unravel the relationships among features (i.e., diagnosis, procedures, medications.). It is reasonable to believe that diagnosis codes, procedure codes, lab values, and medications are usually associated, meaning that they tend to cluster and form patterns. Decoding the relationship among these features is meaningful, which can further help perform patient-level EHR embeddings.

In this work, we address the above two issues by introducing language-based patient embeddings using the EHR data. First, to account for the longitudinal nature of the EHR data, we perform a longitudinal embedding of patients, meaning that for each discrete time point, we will produce a patient vector to represent the specific status of the patient at each time point. We propose to use the year as the discrete unit of EHR events, performing the embeddings that summarize the one-year events of a patient into a vector. This method, on one side, improves the quality of embedding by including the longitudinal feature; on the other side, it provides chances to analyze the progression of phenotypes by examining the changes in longitudinal vectors. Second, to capture meaningful relationships among features, we adopted the autoencoder and transformer architecture to perform the pre-embedding of features and allow features to serve as the basic building blocks for patients. Using our method, we show drastic improvements in the quality of embeddings and demonstrate its great potential for many downstream tasks, including patient topic modelings to identify new patterns, patient bulk phenotyping to classify patients and disease onset predictions.

3.3 Methods

In this section, we describe the detailed model architecture, data, and preprocessing included to build a complete patient embedding model.

3.3.1 Embeddings of diagnosis and procedures

To start, we first perform the embeddings of vocabulary, which are codes of diagnoses and procedures. The embedding of vocabulary is inspired by word representation (or word embedding) [42]. Neural language models usually take embedded words (numerical representations of words) as basic elements for each sentence. We designed an autoencoder neural network architecture for the embedding of diagnoses and procedures, as the goal is to perform a simple compression task of diagnoses and procedures regarding their onset frequency across human life. For instance, across a lifetime of patients, age-related disease would have a distribution of onset frequency skewed towards the late half-life, usually after the age of 60. While pediatric disease or teenage disease would rise early and diminish with time. For each diagnosis and procedure code, we represent their onset frequency in a 1x4320 array, indicating its onset frequency across patients from age 0 to 90, with each cell representing roughly 0.021 years, which equals 7.3 days (thus, roughly a week). To this end, we generated a matrix of 34,851x4320 size, with 34,851 unique diagnosis and procedure codes represented in their onset frequency.

We adopted a simple variational autoencoder neural network to perform the embedding of diagnosis and procedure codes. The model architecture is composed of three layers of dense neural network in the beginning, followed by a mean parameter *z_mean* and variance parameter *z_sigma*. We then stacked the three layers of the dense neural network after the parameter layer to form a symmetric model architecture. This is because the symmetric architecture in autoencoders is known to reduce overfitting, as the number of parameters drop. We selected the size of *z_mean* to be 50, as a good starting point that does not require large computational power but still retains enough information for downstream tasks.

We trained this model with Adam optimizer with default parameters of beta1 and beta2, learning rate 1e-7, and batch size of 64 (Figure 3.1). During the training, 5% of the training data is split and used as a validation set to evaluate the loss. The model is trained for 100 epochs and reaches a steady reconstruction loss and similarity loss. The embedded 50-dimensional codes are then served as vocabularies in the transformer model to construct patient vectors.

3.3.2 Transformer based patient embedding

The embedding of diagnosis and procedure codes in the previous session serves as the basic building block of the patient event. A patient event is defined by all the diagnosis and procedure codes sequentially in a certain year for a patient. Therefore, for individual patients, there could be multiple patient events, as long as the EHR contains patient visits across different years. In this case, we perform embedding of each patient event, using the transformer model architecture. Like any language models, the transformer model takes sentences that are represented by embedded vocabularies as input. In this work, each patient event can be seen as an independent sentence, with embedded diagnosis and procedure codes as vocabularies. The traditional transformer model architecture can be seen in [60,68]. Here we adopted the basic model architecture and represented the model in Figure 3.1. For this task, we did not include the positional embeddings for each code, as we binned all codes for each patient happening within a year into a single vector, we do not meticulously require a specific positional encoding of codes within a specific year. We use the hyperparameter L = 6, H = 10, diff = 2048, and d = 200. The pre-trained embeddings of codes from the autoencoder serve as vocabularies. In the preprocessing step, we bin each patient's codes by year, and for each year, we create a vector of codes to represent the temporal patient vector. The maximum vector length is set to be 250, considering that in one year, most patients should receive less than 250 codes from the hospital. For all 102,740 available patients, only 0.25% have more codes than 250 in a year, which is rare and might only happen to extremely severe patients. For patients having codes less than 250, we apply zero padding to the sequence and mark it as a special padding token. In total, there are 1,046,649 patient

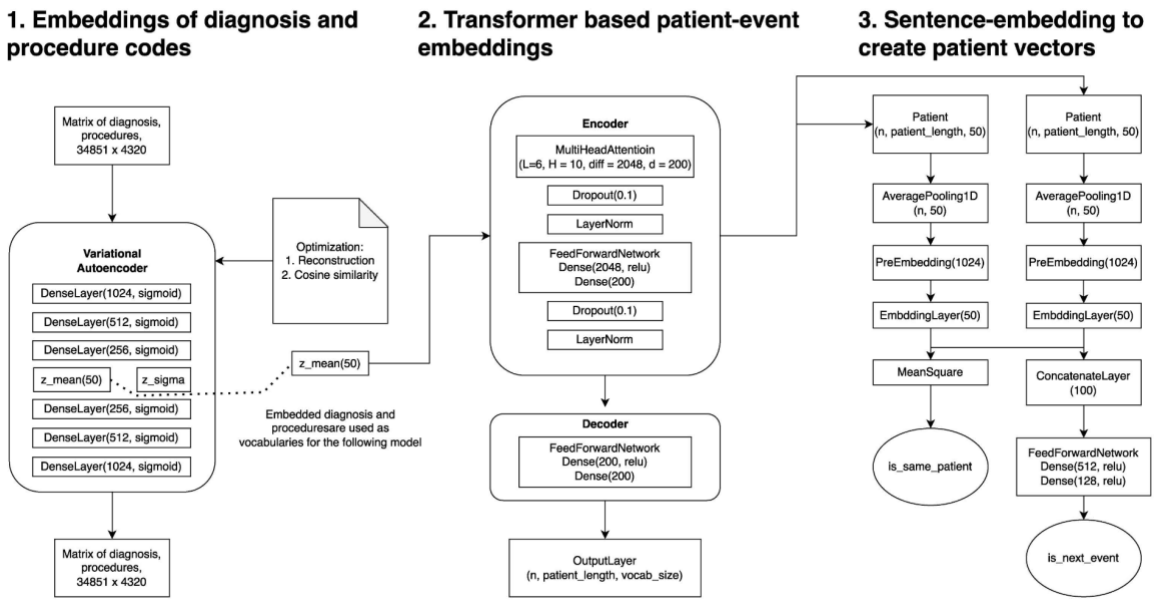
vectors for 102,740 patients. The goal of this transformer model is to reconstruct the codes of patients. During the training process, we masked 20% of the codes in each vector and let the model reconstruct the full sequence of vectors. We trained this model with Adam optimizer with beta1 = 0.9, beta2 = 0.98, epsilon = 1e-9, and a scheduled learning rate gradually decreased to 0.002 at step 10,000. The batch size is set to 32, reducing the computation memory load. Again, 5% of the data is split and used as validation to evaluate the loss during the training.

3.3.3 Sentence-BERT based patient embedding

As the output of the transformer model is a multi-dimensional vector, instead of a one-dimensional representation of arrays, we adopted the sentence-BERT (SBERT) based model to perform the dimensional reduction that is specifically tuned on two downstream tasks. Similar to the S-BERT model, we also built our model architecture that can take two patients as input while giving two binary outputs. The first binary output of our model denotes whether the two vectors of patients belong to the same patient (*is_same_patient*). The second output predicts if one patient event is chronologically the following event of another patient event (*is_next_event* task). The detailed model architecture is included in Figure 3.1. For this task, we first build a global average pool, taking the embedding layer from the transformer model as input (Figure 3.1). Note that we feed two embedded patient events at one time for this model. We then added a feed-forward structure of a pre-embedding layer and a 50-dimensional embedding layer to compute the complex interactions of the embedded sequence (Figure 3.1). We use two different loss functions to optimize the two tasks mentioned above (*is_same_patient* and *is_next_event*). The goal of the *is_same_patient* task is to minimize the mean square distance between two 50-dimensional embedded vectors, as an evaluation of the vector distance in geometric space. While the *is_next_event* task is optimized by a 2-layer feed-forward neural network constructed by concatenating the embedding of two vectors, this might allow learning of complicated relationships among two vectors. We only used two layers of feed-forward structure to avoid overfitting and to ensure it learned meaningful functions according to the theory that two layers of neural networks can simulate any form of continuous function. During the training, we

randomly formed pairs of patient vectors as input and trained until no improvement (loss decrease). We repeated the training process a few times, considering the variation due to the randomness of parameter initialization. We found the model performance peaked at 5000 steps during training with a learning rate = 3e-4.

Figure 3.1 Illustration of the model architecture



3.3.4 Data

Patient EHR data from the eMERGE Network (n = 102,740) were used for training and building the patient embedding models. These included basic demographics (birth decade, gender, race, and ethnicity), patient diagnosis codes (ICD-10 and ICD-9), procedure codes (CPT-4), and age at diagnosis. The UW EHR data (n = 840,000 available patients) served as a validation set. Besides the same data elements mentioned above (birth decade, gender, race, ethnicity, diagnoses, and procedures), UW included hospital mortality data, which is then used to evaluate the survival differences among clusters.

3.3.5 T-distributed stochastic neighbor embedding (TSNE)

TSNE, which stands for t-Distributed Stochastic Neighbor Embedding, is a machine learning algorithm used for dimensionality reduction and visualization of high-dimensional data. Developed by Laurens van der Maaten and Geoffrey Hinton, t-SNE is especially effective at projecting complex data structures into low-dimensional spaces while preserving local relationships between data points [70].

TSNE works a little bit differently than the traditional matrix decomposition method, which oftentimes uses a linear combination of features (such as PCA). TSNE took a probability approach, which calculates the probability of similarity between points using a Gaussian distribution in high-dimensional space and a Student's t-distribution in low-dimensional space (visualization space). Then, TSNE iteratively optimizes the layout of points in the lower-dimensional space by minimizing the Kullback-Leibler divergence between the high-dimensional similarity distribution and the low-dimensional similarity distribution.

3.3.6 Softwares, versions and code availability

The patient embedding model is implemented through tensorflow version 2.3.0, with mostly the high-level API tensorflow.keras, version 2.4.0. Models were trained using an NVIDIA 2060-Super GPU with 8 GB RAM. The code for running the model and synthetic data are available at the GitHub repo https://github.com/suxian06/language-model-based-patient-embedding/tree/main. Data analysis using TSNE, PCA, GMM, and BIC was implemented through the scikit-learn version package 0.24.2 in Python. Logistic regression and analysis of variance (ANOVA) were implemented using statsmodels, version 0.12.2. Plots were generated using Matplotlib version 3.4.3, Seaborn 0.11.2. The online interactive charts were generated using Altair version 5.0.1. Survival analyses (Kaplan-Meier plot and Log-rank test) were performed using the scikit-survival packages in Python, version 0.14.0. Log-rank test used in differentiating the subgroup survival were also based on the scikit-survival package (compare_survival function).

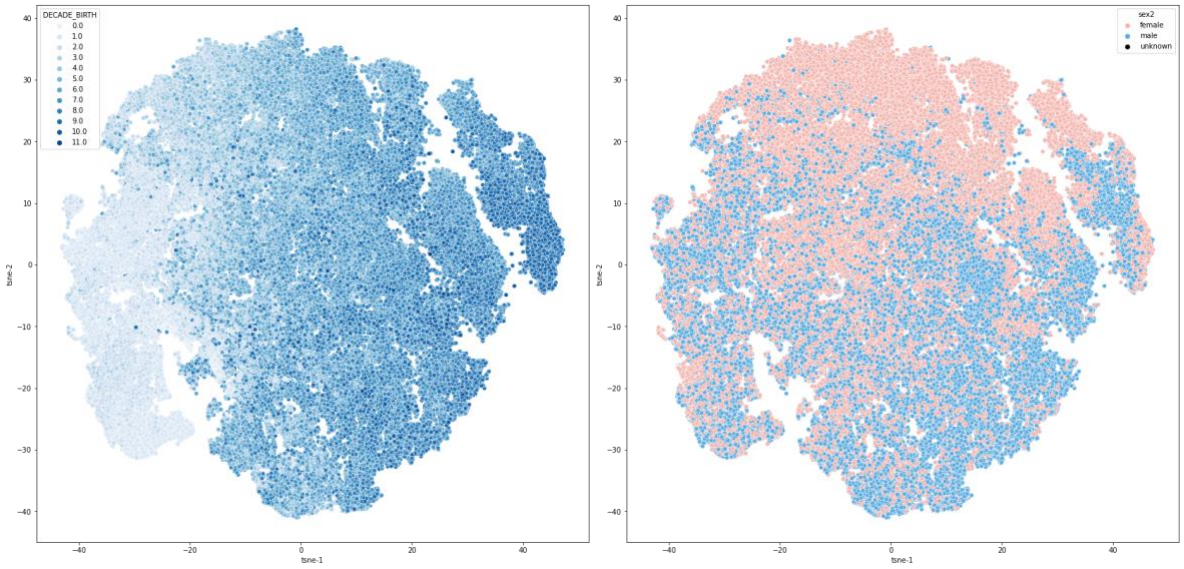
Standardization, numerical operations, and data cleaning are done with numpy version 1.23.0 and scipy version 1.6.2.

3.4 Results

3.4.1 Model performance evaluation

After the three model components, each patient event is represented by a 1x50-dimensional array. The embedded patient event vectors for a single patient are a numerical representation of all the diagnosis and procedure information of a patient, with partial information about their relative onset time throughout the life of the patient. We embedded all the patients from the eMERGE Network into this numerical space and aimed to perform downstream tasks in the following chapters. To start, we represented all embedded patients in a two-dimensional space using T-distributed stochastic neighbor embedding (TSNE, Figure 3.2). In Figure 3.2, we observed that the axis of TSNE-1 is driven by patient age, while the axis of TSNE-2 represents great gender differences. The results pointed out two major factors in all the modern analyses, as age and sex are always included in various clinical models as covariates.

Figure 3.2 Two-dimensional TSNE representation of patient embedding. Left panel is colored by patient age (the darker the older) and the right panel is colored by patient gender (pink to female, blue to male)



To quantitatively evaluate the model performance and understand how well the information is preserved, we performed two tasks to ensure the quality of the embedding. First, to evaluate how well the information is preserved, we feed the embedded patient event vector into the transformer model (the second component of the three-step model architecture) and evaluate whether the embedded patient event vector can recover the original diagnosis and procedure codes. Second, to ensure the embedded patient vector tuned in the last model step can distinguish different patients (*is_same_patient*) and recognize the next event chronologically (*is_next_event*), we performed a random selection of patient events and evaluated the performance on both tasks. The result is presented in Table 3.1.

Table 3.1 Performance of reconstructing original diagnosis and procedure codes.

	median	mean
precision	0.958	0.917

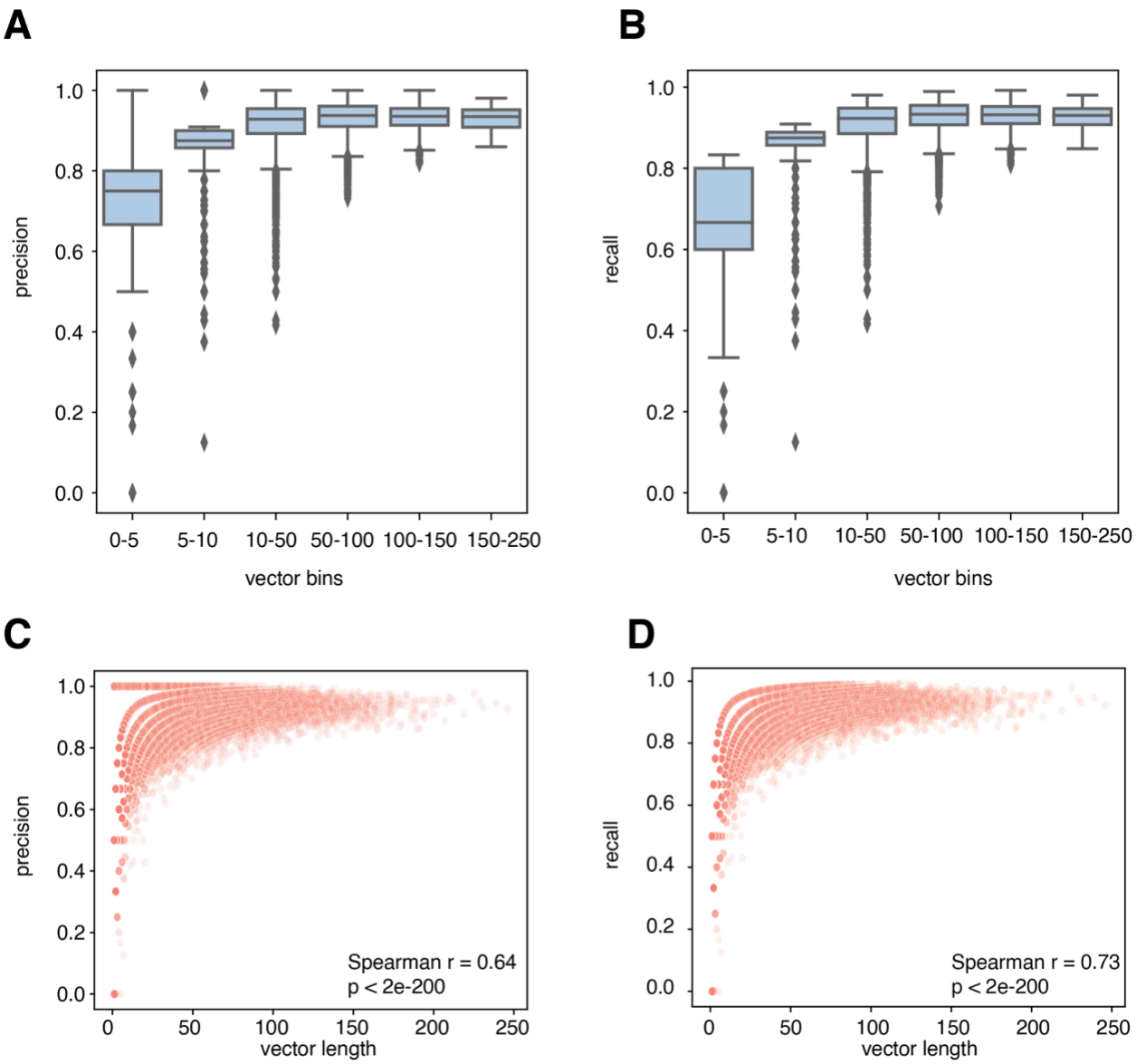
recall	0.937	0.895

The model achieved incredible results on the reconstructing original diagnosis and procedure codes task. As it is known that the attention mechanism can capture feature relationships very well, it is no surprise that both median and mean precision and recall are almost above 0.9. However, as the attention mechanism was originally designed for pairwise language translation, the length of the sentence, for here, is the length of the patient event, might influence the performance. As hypothesized, we did observe a performance difference in the length of the patient event, represented by the number of codes within each patient event (Figure 3.3)

Figure 3.3 Performance (precision and recall) differences in patient events with different numbers of codes.

A-B. Precision (A) and recall (B) based on the numbers of codes binned into ranges in each patient event (x-axis).

C-D. Same relational plot of A-B but in continuous manner without binning numbers of codes into ranges.



For the S-BERT model performance evaluation, we randomly selected 500 patients to form pairs within 5 iterations, to evaluate the performance on *is_same_patient* and *is_next_event*. In total, we evaluated 29,587 events within the eMERGE dataset. The detailed results are included in Table 3.2. The model achieved reasonable results, with an accuracy of 0.797±0.0016 for *is_same_patient* and 0.769±0.011 for *is_next_event*. Note that 500 patients forming random pairs might not sound representative enough, but the number of events evaluated (n = 29,587) is not trivial. Moreover, the standard deviation of performance within each iteration here is extremely low, indicating a robust performance.

Table 3.2 Evaluation of the *is_same_patient* and the *is_next_event* performance for the sentence-embedding model

	is_same_patient	is_next_event
Accuracy	0.797±0.0016	0.769±0.011

3.4.2 External validation using local UW EHR data

One of the common barriers and limitations of deploying machine learning mode is that most models are evaluated internally, without external validation. To address this issue, we collected the EHR data at the University of Washington (UW) from 2000 to 2020, including n = 840,000 patients as external sources of validation to assess our model's validity and robustness. We will evaluate the models based on three of the downstream tasks, including high-throughput phenotyping, disease onset prediction, and heterogeneity analysis.

First, albeit not comparable to the performances on the eMERGE dataset (Table 3.1), the transformer model trained on the eMERGE data still earned a reasonable performance in sequence recovery tasks running on UW patients, with a mean and median precision of 0.862 and 0.903, respectively. The mean and median recall are 0.852 and 0.896 (Table 3.3). Next, the performance of the sentence embedding model has an accuracy of 0.796±0.0023 in the *is_same_patient* task and 0.742±0.0042 for the *is_next_event* task on the UW dataset (Table 3.4). This result is interesting as it is comparable to the performance on the training set (eMERGE), and we can conclude that the performance is ideal for external validation (Table 3.4). Together, this evidence indicates that the transformer model can achieve great success in patient embedding internally and is also generalizable with the external UW cohort.

Table 3.3 External validation of the is_same_patient and the is_next_event performance for the sentence-embedding model using the UW cohort

	median	mean
precision	0.903	0.862
recall	0.896	0.852

Table 3.4 Evaluation of the is_same_patient and the is_next_event performance for the sentence-embedding model using the UW cohort

	is_same_patient	is_next_event
Accuracy	0.796±0.0023	0.742±0.0042

3.5 Conclusion

In conclusion, to study complex patterns within diseases, we designed this language-model-based patient embedding method, representing patient information from the EHR into numerical space. This is the first step in our exploration of the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. This step serves as a starting point for diving deep into diseases of interest and patient stratifications, which we will demonstrate in the following chapters (Chapter 4-6).

In this chapter, aligned with the goal in Aim 1 (to develop patient representation learning in EHR data using an unsupervised machine learning approach), we developed a novel patient embedding method working with the EHR data. We integrated 3-step model architectures to achieve this complicated task and

demonstrated a few downstream applications. Qualitatively, we showed that the embeddings can reveal gender and age differences on the TNSE-1 and TSNE-2 axis. Quantitatively, we evaluated the model performance both internally and externally. We demonstrated that the model, especially the last two components, can accurately reconstruct information from the embeddings (Table 3.1 and Table 3.2). Though the external model performance on the UW EHR data dropped, the metrics score (Table 3.3 and Table 3.4) is still satisfying, experimentally demonstrating the robustness of our model. An interesting point is that we did notice a drastic difference in the model performance on various lengths of patient vectors (Figure 3.3). This variation is related to the attention mechanism. The attention mechanism is designed for pairwise translation, where a sentence usually consists of a proper number of words. Therefore, when a patient sequence sparsely contains very few codes, the model lacks a strong co-occurrence pattern to recover the original codes.

We noticed a trend of more complicated model architectures and downstream applications within the development of this new area, indicating a promising future for EHR representation learning. For example, other studies have explored the potential of representation learning to characterize diseases using the EHR with various machine learning architectures [11,12,14,53,71]. Most of these studies use different metrics and disease annotations, making it challenging to compare the performances side-by-side. Among them, our model used a novel vocabulary embedding strategy to represent the diagnosis and procedure codes in the onset frequency domain. Our model showed more effortless components of embedding (only diagnosis and procedure codes and high computational efficiency) but still achieved vigorous performances. To our knowledge, we are the first group using a complex source of EHR data across 12 sites to perform the representation learning, leading to a thorough and more generalized patient representation model. Additionally, we seem to be the only group that examined the model performance externally, validating the results using a local UW EHR dataset.

Mining information from the EHR has become a crucial topic for several advanced downstream implementations, such as disease prediction, patient phenotyping, and personalized medicine [19,72]. In the following chapter, we will examine several crucial downstream tasks and demonstrate the use cases of the patient embeddings.

Chapter 4: High-throughput Phenotyping

4.1 Overview

This chapter is the second step of our explorations of the secondary use of EHR from both unsupervised and supervised methods, which leads to the utilization of the EHR data to identify novel disease patterns. In this chapter and the following chapter (Chapter 5), we focus on a demonstration of patient embedding applications in classic clinical tasks. These two chapters (Chapter 5 and Chapter 6) bridge Aim 1 (to develop patient representation learning in EHR data using an unsupervised machine learning approach) and Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors), serving as a prelude and quality assessment of traditional clinical tasks before transitioning to the exploratory analysis of novel disease patterns in Chapter 6. Here, we start with phenotyping in this chapter, as one of the crucial cohort-building methods, showing how patient embedding enables automated high-throughput phenotyping.

With the wide adoption of the EHR system in the United States, EHR phenotyping has become one of the most crucial and standard research topics within this field.

EHR phenotyping, by definition, is the use of EHR data to digitally define if a patient has a specific trait of interest or not. EHR phenotyping uses available information from the EHR system, including diagnosis, procedures, labs, medications, observations, and sometimes clinical notes. EHR phenotyping has many promising downstream applications, including facilitating genome-wide association study (GWAS) using EHR-defined phenotypes and performing clinical trials with EHR-defined cohorts.

The detailed use of EHR data sometimes relies on existing knowledge and domain expertise, as commonly seen in manual or rule-based phenotyping algorithms. The manual or rule-based phenotyping approach has

achieved great success in the field. The eMERGE network, combines efforts across more than ten clinical sites, gathering data and publishing more than 60 phenotyping algorithms, stored in pheKB.

However, the manual or rule-based phenotyping approach is heavily laborious and known to be iterative in development. Recently, a growing number of machine learning approaches appeared within the community aiming to reduce the laborious and potentially boost performance. Machine learning algorithms use the same EHR data and incorporate both supervised and unsupervised approaches to dealing with phenotyping tasks. Though some studies have shown some promising performance using machine learning, no significant performance boost compared to manual or rule-based algorithms is found.

4.2 Related Work

High-throughput phenotyping can largely accelerate clinical and translational research using data from the EHR. Nowadays, efforts of high-throughput phenotyping can be divided into rule-based and machine-learning-based. Rule-based methods are developed through iterative processes with large manual effort from expertise with clinical knowledge. One of the successful examples of rule-based high-throughput phenotyping is phecodes, which directly maps ICD codes to about 1800 unique phenotypes, and demonstrates its use in phenome-wide association study (PheWAS) [73]. The machine-learning-based approach is more commonly used in high-throughput phenotyping, due to the superior ability in identifying complex patterns. Several approaches have been reported, including mostly supervised and semi-supervised approaches. For example, PheCAP is a semi-supervised approach that utilizes structured and unstructured data processed by natural language processing techniques to perform phenotyping [74]. Another similar approach uses active learning to demonstrate improvement in phenotyping tasks [75]. Though increasingly complicated approaches are taken to address these issues, the major focus of high-throughput phenotyping is to provide an automated framework for labeling participants [76]. The biggest concern for these tasks is that the gold-standard definition of a phenotype and data sources vary, making the direct comparison of algorithms extremely challenging.

4.3 Methods

4.3.1 Compute mean vector for each patient.

High-throughput phenotyping is a strategy to output a wide range of possible phenotypes for an individual. In this work, we perform the high-throughput phenotyping using the embedded patient vector as features. For each patient (*patient*$_k$), we first computed the mean of vectors (*meanVector*$_k$) across all time points (t) within individual patients (*patient*$_k$), resulting in a single vector for each patient.

$$meanVector_k = \frac{\sum_{(t=1...n)} patientVector_k^t}{n}$$

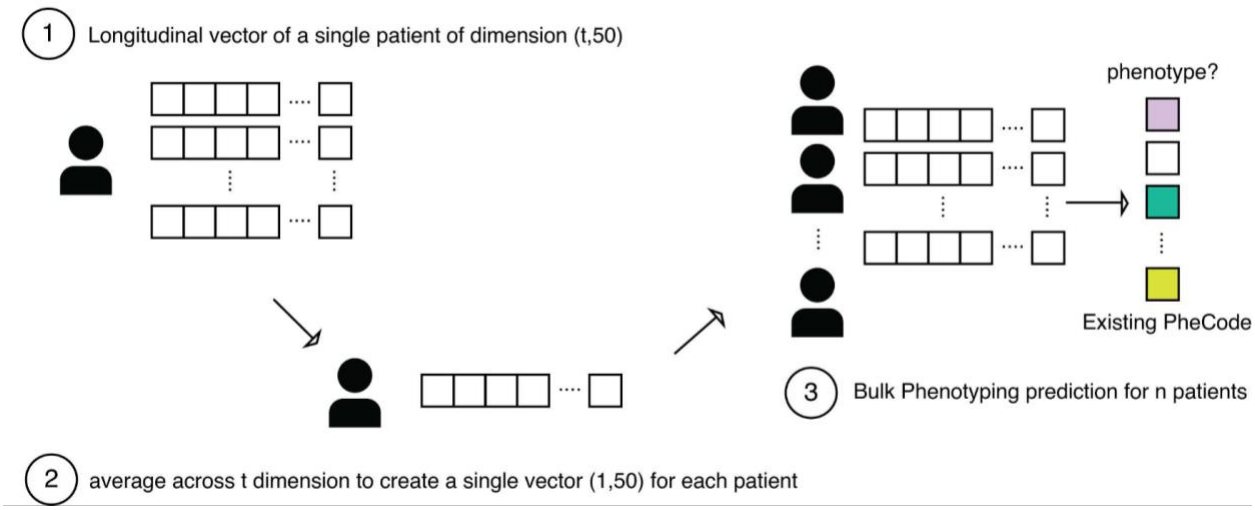
4.3.2 Build a logistic regression model for each phenotype

Then for each phenotype (*phenotype*$_i$) defined using phecodes, we encoded it into binary representation. If a patient has *phenotype*$_i$, we encode the outcome to be 1, otherwise 0. Then for each phenotype *phenotype*$_i$, We build logistic regression models using these vectors to discern whether patients exhibit a specific phenotype or not, as defined by phecodes. For each model and each phenotype, 20% of the data is used to evaluate the performance (test set) and 80% of the data is used for training.

$$phenotype_i = \frac{1}{1 + e^{\sum -\beta_j x_j}}$$

The step-by-step illustration is in Figure 4.1.

Figure 4.1 Step-by-step illustration of high-throughput phenotyping.



4.3.2 Mapping ICD10 and ICD9 codes to phecode

The mapping between ICD9 and ICD10 diagnosis codes to phecodes is done by referencing the PheWas website https://phewascatalog.org/ through the phecodes mapping panel, counting at least one presence of diagnosis code as qualifying phenotypes. This step is critical to remove redundancy and too detailed granularity annotated by ICD9 and ICD10 codes. After mapping, we obtained 1855 unique phenotypes defined by phecodes.

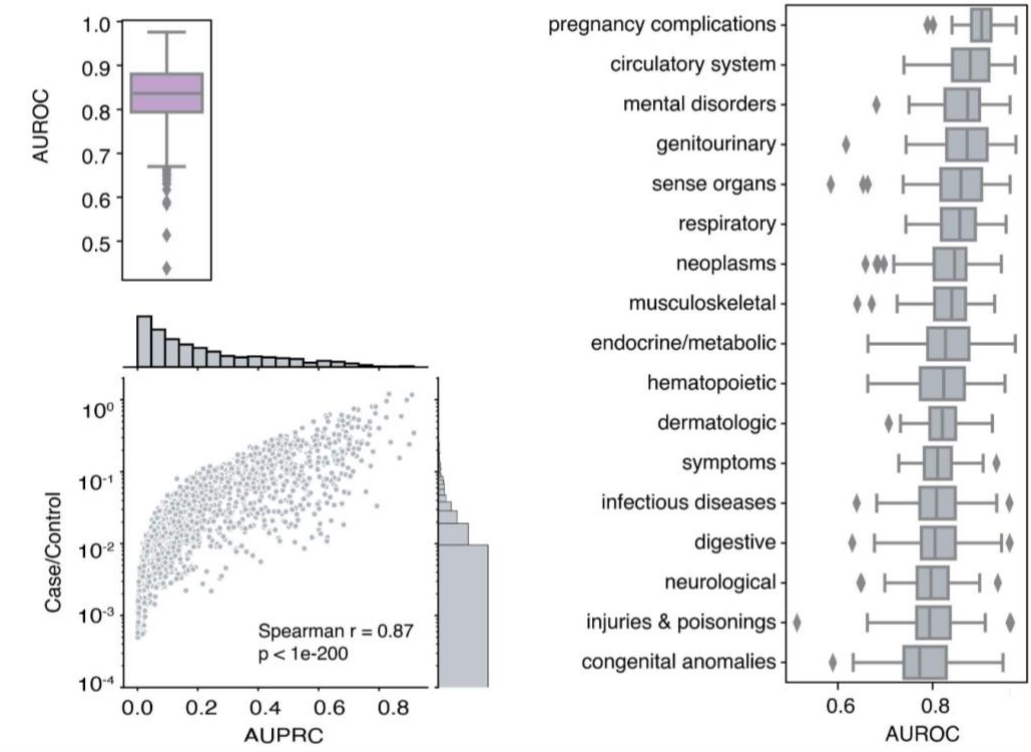
4.4 Results

4.4.1 High-throughput phenotyping using eMERGE data

For the high-throughput phenotyping task, we report the performance in general and also illustrate the performance by disease categories. As mentioned above 80% of the data is used for training, and the evaluation is performed on 20% of the test set. In general, we made predictions on 1855 distinct phenotypes, defined by PheCode and achieved a good performance. The mean area under the receiver operating curve (AUROC) for the 1855 distinct phenotypes is 0.84, indicating a good model fit (Figure 4.2). Moreover, when breaking down by disease categories, the performance in each disease category is similar, with the

best AUROC = 0.90 for pregnancy complications and the least AUROC = 0.77 for congenital abnormalities. These results together, suggest the patient embedding is robust and not biased towards any disease categories and can achieve great performance in high-throughput phenotyping task.

Figure 4.2 Performances on high-throughput phenotyping. Top right showing the boxplot of AUROC and bottom showing the relationship between sample size (case/control ratio) and AUPRC. On the right each boxplot represents AUROC distribution categorized by disease class according to phecodes.



4.4.2 External evaluation using local UW EHR data

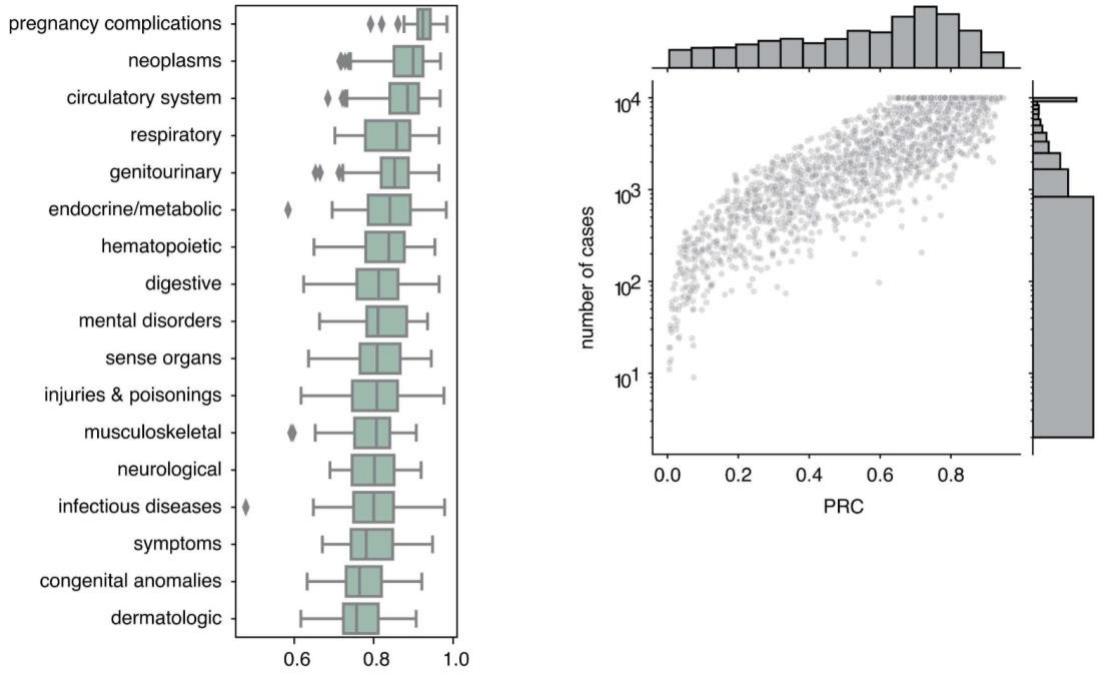
To achieve a more unbiased quality assessment of the trained model, we applied the model, trained using eMERGE data, to data extracted from the local UW EHR and evaluated its performance on high-throughput

phenotyping. The UW data comprises 840,000 patients, spanning from 2000 to 2020 (3.3.4 Data). Although the data collection periods for UW and eMERGE are nearly identical, there are some discrepancies. First, the eMERGE data includes all lifetime diagnosis and procedure codes for individuals, providing a more complete reflection of medical history, whereas the UW EHR data only covers up to 20 years of history per individual. Additionally, the UW data, sourced from a West Coast population, likely differs in environmental, behavioral, and cultural factors, which can influence medical patterns. Nevertheless, we believe that if our model is trained with a representative sample of patients, it should demonstrate robust performance on the UW dataset, despite these potential population variations.

For the evaluation step, we also applied the same strategy, building a logistic regression model for each phenotype and using 80% for training while retaining 20% for evaluation (4.3.2 Build a logistic regression model for each phenotype).

For the evaluation step, we applied the same strategy by building a logistic regression model for each phenotype, using 80% of the data for training and retaining 20% for evaluation (4.3.2 Build a logistic regression model for each phenotype). Detailed results are presented in Figure 4.3. The model achieved a median AUROC of 0.83 for the high-throughput phenotyping tasks, only 0.01 lower than the performance on the eMERGE dataset. This result is highly consistent with the eMERGE data, with the highest AUROC observed for pregnancy complications, while congenital anomalies ranked as the second lowest. The performance is robust and surprisingly stable for an external evaluation.

Figure 4.3 External evaluation of the performances on disease onset prediction. The left panel shows the boxplot of AUROC categorized by disease class according to phecodes. The left panel shows the relationship between sample size (case/control ratio) and AUPRC.



4.5 Conclusion

In this chapter, we explored the ability of patient embeddings for high-throughput phenotyping, addressing the objective of Aim 1 (to develop patient representation learning in EHR data using an unsupervised machine learning approach). We demonstrate that a simple linear combination of the embedded features can manage to get a good performance on high-throughput phenotyping tasks. In this chapter, we explored the secondary use of EHR from an unsupervised method, exploring the potential of patient embedding utilizing the EHR data to identify novel disease patterns.

As the embedding (the mean vector) conceptually represents the trajectory of a patient in the EHR, high-throughput phenotyping is similar to the extraction of information at a specific time point. Thus, we need to be aware that for particular patients, there is a potential for a dominant phenotype across the trajectory that hinders the granularity and performance of high-throughput phenotyping.

One interesting fact is that this method is data-driven and thus not limited to existing knowledge for EHR-based risk prediction and can uncover new disease patterns. Some studies even use multiple models and allow paralyzation learning to perform semi-supervised learning, taking advantage of the data-driven method.

An important aspect we noticed is that most machine learning applications have not been robustly evaluated externally. In this session, we utilized the UW cohort from 2000 to 2020 to perform an external evaluation. Even though the data source, sample sizes, and the distribution of codes are all different (the UW cohort only spans 20 years, while eMERGE cohort extracts lifetime codes for individual participants), the external performance is great, demonstrating the robustness of our model.

In the subsequent chapter (Chapter 5), we will apply patient embeddings in another classic clinical task -- disease onset prediction, following a similar format to the current chapter.

Chapter 5: Disease Onset Prediction

5.1 Overview

This chapter, similar to the previous one, focused on a traditional clinical task -- disease onset prediction -- before transitioning to the exploratory analysis of novel disease patterns in Chapter 6. These two chapters together provided illustrations of patient embedding on canonical clinical tasks with thorough quantitative assessment and validations, designating the fulfillment of Aim 1 (to develop patient representation learning in EHR data using an unsupervised machine learning approach) and establishing a solid ground for the road to Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors). In this chapter, we continue the exploration of the secondary use of EHR from an unsupervised approach, which paved the way for utilizing EHR data to identify novel disease patterns.

Forecasting diseases in the future is always a crucial task, as early diagnosis of disease usually means on-time treatment and better clinical outcomes. In the current era, disease diagnosis relies on clinical symptoms, observations, lab tests, and sometimes radiology findings. With the advancement of genomic technology, researchers are building novel risk prediction tools based on genetics, such as polygenic risk scores, to predict the risk of getting a specific disease. Modern genetics has proved great power in identifying high-risk patients and providing more effective prevention. However, genetic risk predictions are not always sufficient, especially for complex diseases that arise from interactions between lifestyle, environment, and genetics. For example, the AUROC of polygenic risk score on colorectal cancer and many other complex traits can barely surpass 0.6, indicating that the variance is not solely captured by genetics.

One of the most crucial tasks in the clinic is to predict whether a disease will occur or not in the future, given the current information. Several works started to use complicated EHR data to forecast the future. Reasonably, the sign of the onset of a future disease might already be hidden within the current information

system. Using the currently available information from the EHR to predict the future is feasible at a certain point.

In this work, we will use the embedded patient vectors are features to perform disease onset prediction tasks for 1855 distinct phenotypes defined by PheCode. The intuition behind this is that certain diseases are enriched with specific comorbidities. This means that if the embedded patient vector captures the co-occurrence pattern of diseases, we would be able to predict the short future given the current information.

5.2 Related Work

Disease onset prediction is a common and crucial task, which can be rephrased as risk prediction. There are many methods for disease onset/risk prediction. Nowadays, using genetic data to predict disease risk has become very popular. It is common and accurate to use genetic markers for Mendelian disease. However, the performance drops when predicting complex diseases/traits that are not monogenic, even with the help of polygenic risk scores.

In the current era, EHR data contains rich information for individuals in a longitudinal manner, which can be helpful for future disease predictions. The Deep Patient has utilized almost all data types from the EHR to perform patient embeddings and further applied the embeddings to predict future diseases [11].

Med-BERT, pre-trained embedded medical vocabularies, is also used for disease predictions under the context of EHR. This work is conceptually refreshing, as it leverages medical vocabularies in a semantic way to perform disease predictions [53].

Several other works utilize the semantics of medical vocabularies to build language models that can perform disease prediction tasks [65,77,78]. In this case, our models are embedded based on the medical code frequency, instead of semantic vocabularies, which are the core differences from others. We believe that

the onset frequency of codes can better capture the longitudinal pattern, thus providing better performance in disease onset predictions.

5.3 Methods

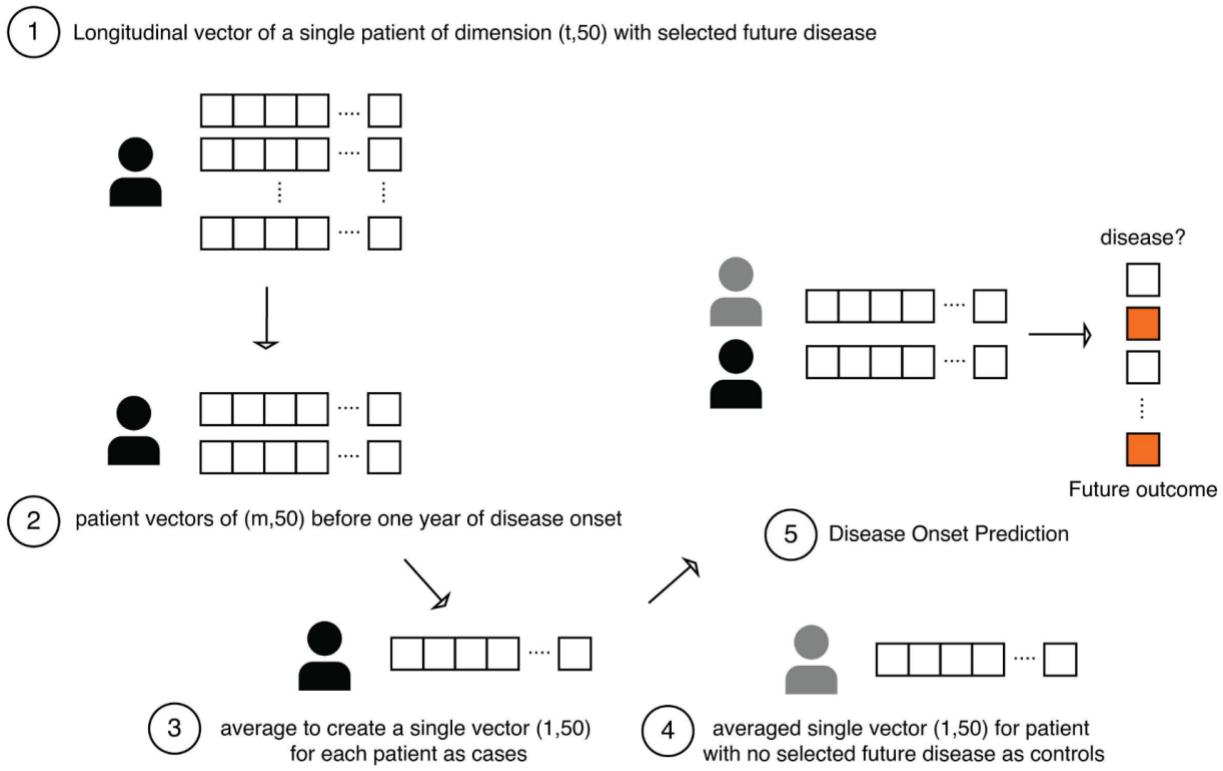
5.3.1 Build logistic regression model for each phenotype.

Both disease onset prediction and bulk phenotyping are classification tasks in this work. In disease onset prediction, for each phenotype, we collected all longitudinal vectors of patients and split them into before the onset versus after the onset group. Then, the concept is very similar to high-throughput phenotyping work, where we used the longitudinal vectors to build a logistic regression classifier to perform classification tasks for the two groups of vectors. For example, for each disease *i*, for *j* in 1…50, x_j represents numerical features drawn from the embedding (embedding size of 50), the logistic regression prediction whether the disease *i* is already presented in the longitudinal vector or not:

$$onset_i = \frac{1}{1 + e^{\sum -\beta_j x_j}}$$

The step-by-step illustration is included in Figure 5.1. We also applied logistic regression for the sensitivity analysis with different years as onset thresholds. Instead of splitting vectors into two groups, we only included the longitudinal vectors that are 1 year, 3 years, or 5 years before the disease onset, to perform the prediction task.

Figure 5.1 Step-by-step illustration of disease onset prediction.



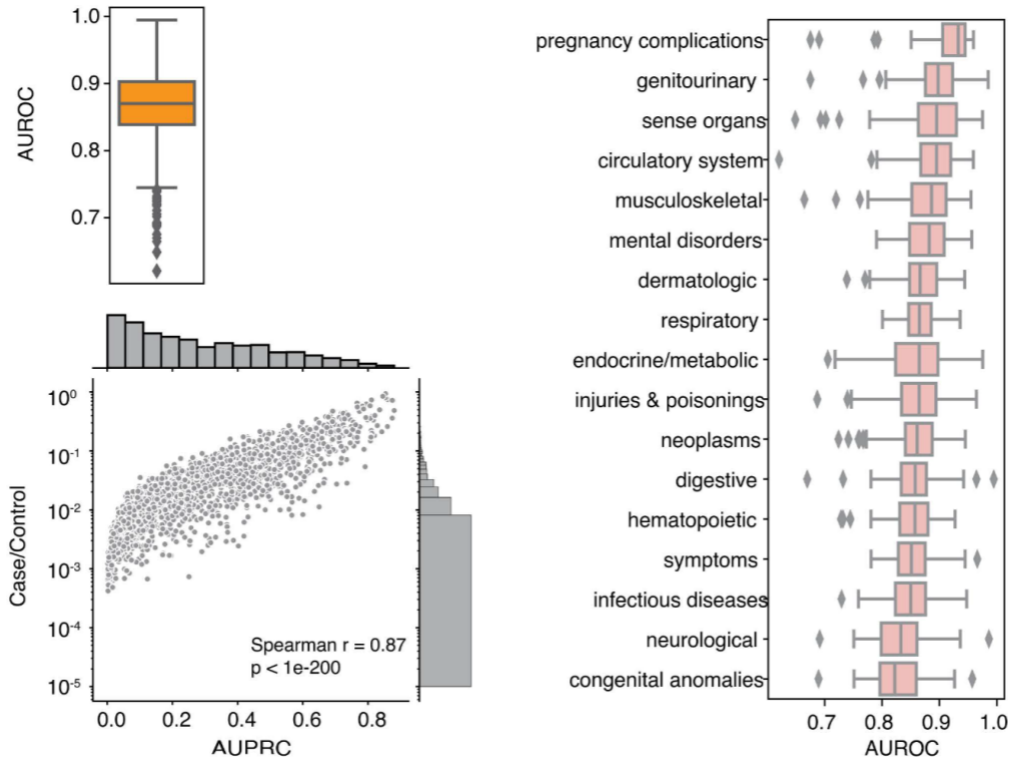
5.4 Results

5.4.1 Disease onset prediction performance on eMERGE data

Disease onset prediction, as seen in the above method session, is similar to the high-throughput phenotyping task. However, the disease onset prediction is conceptually more challenging and clinically more critical. Predicting future disease onset requires prior knowledge and baseline risk assessment of individuals, with the potential of accumulating risks during aging. In this session, we will demonstrate the use of patient embedding to predict future disease onset. The motivation and foundation behind this is that the patient embeddings are learned longitudinally with the diagnosis and procedure vocabulary embedded in the onset-frequency domain. We believe this would boost the performance of disease onset prediction.

For this task, we again report the performance in general and also illustrate the performance by disease categories, similar to what we have done in high-throughput phenotyping. As a standard, we used 80% of the data for training and then evaluated 20% of the untouched test set. The mean AUROC for the 1855 distinct phenotypes is 0.87, indicating a good model fit (Figure 5.2). For each disease category, the best AUROC is 0.93, for pregnancy complications and the least AUROC is 0.82 for congenital abnormalities. Consistently, the performance in each disease category is smooth and even, without any drastic differences.

Figure 5.2 Performances on disease onset prediction. Top right showing the boxplot of AUROC and bottom showing the relationship between sample size (case/control ratio) and AUPRC. On the right each boxplot represents AUROC distribution categorized by disease class according to PheCodes.



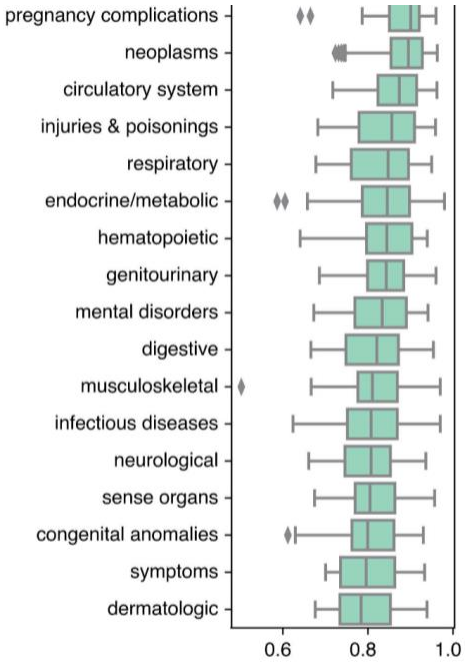
Considering adding sensitivity analysis, and distinguishing results.

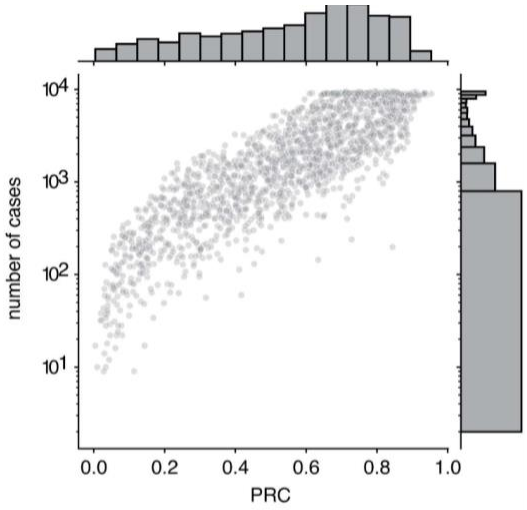
5.4.2 External evaluation using UW local EHR data

The external evaluation session of the disease onset prediction is similar to the high-throughput phenotyping task. In a short summary, we build individual regression models for each phenotype and use 80% as the training set while evaluating the 20% hold-out set (5.3.1 Build logistic regression model for each phenotype).

The detailed performance is summarized in Figure 5.3. We achieved a median AUROC of 0.84 on disease onset tasks, across all 1855 phenotypes, with only 0.003 lower than the performance on the eMERGE dataset. This tiny performance drop in the UW datasets is reasonable, as the model usually over-fit the training data at a certain point. Still, the performance is robust and surprisingly stable for an external evaluation. The evaluation task here successfully demonstrated the great performance of the model on an external dataset. Additionally, the performance on the external dataset is stable across all categories of phenotypes, similar to the training data, showing the steady performance of our model.

Figure 5.3 External evaluation of the performances on disease onset prediction. The left panel shows the boxplot of AUROC categorized by disease class according to phecodes. The left panel shows the relationship between sample size (case/control ratio) and AUPRC.





5.5 Conclusion

This and the previous chapter together, fulfilled the goal of Aim 1 (to develop patient representation learning in EHR data using an unsupervised machine learning approach), with a thorough demonstration of phenotyping and disease onset prediction tasks, serving as an excellent quality assessment of patient embeddings. These two chapters both focused on tasks that can be quantitatively evaluated, where solid and robust performance ensures the quality of patient embeddings. In Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors), shown in the following chapter (Chapter 6), we will perform exploratory analysis and evaluate the results based on existing knowledge. By now, we explored the secondary use of EHR from an unsupervised approach, which will lead the way for identifying novel disease patterns.

Disease onset prediction is crucial for several reasons. First, it allows on-time detection or risk assessment to initiate early diagnosis and treatment, which is known to be beneficial for many conditions. Second,

automation in disease onset prediction can largely decrease the burden of physicians and establish a better healthcare system for using artificial intelligence.

As stated before, the embedding (the mean vector) conceptually represents the trajectory of a patient in the EHR. Plus, the transformer model is specifically tuned to predict the next year's code given the current year. Therefore, we expect the model to excel, especially in disease onset prediction tasks.

As we anticipated, the model reached a slightly better performance compared to high-throughput phenotyping. The reason is stated above, as we specifically tuned the model to be good at capturing the patient's longitudinal trajectory.

During the evaluation of the UW cohort, we observed reasonably good performance overall. As mentioned before, the UW cohort is drastically different from the eMERGE cohort (the training cohort) in almost every aspect. Thus, the results demonstrated the external robustness of our model, indicating great potential for deployment.

This chapter, together with the previous chapter, meets the objective of Aim 1 (To develop patient representation learning in EHR data using an unsupervised machine learning approach) and provides a thorough model assessment. In the following chapter, we will dive deep into novel disease patterns explorations using patient embeddings.

Chapter 6: Heterogeneity Analysis in Disease Progression

6.1 Overview

In previous chapters, we produced patient embeddings based on our novel language-model-based method. Here, we dive deep into uncovering novel disease patterns based on the heterogeneity of patients within certain diseases of interest. The results from previous chapters consolidated our confidence in the embedding quality, especially with demonstrations of two traditional clinical tasks, phenotyping, and disease onset prediction. In this chapter, our goal aligned with Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors), focusing on studying the heterogeneity of patients within a predefined phenotype (here using CRC and SLE as two examples) and investigating progressional discrepancies. This chapter is the last step we took, exploring the secondary use of EHR from the unsupervised approach and uncovering novel disease patterns within the EHR data. While the next chapter will mainly focus on the study of disease etiology.

Patients are all different in all aspects. For example, depression, as a psychiatric disease, can have different effects on patients depending on the circumstances. Sometimes, depression lasts around days and patients recover soon. In severe cases, clinical depression can even cause patients to hurt themselves or commit suicide. Generally, this rule applies in all diseases and phenotypes, that each phenotype or disease should not be considered as a simple binary measurement.

The above is known as internal heterogeneity for a single specific phenotype. More specifically, not only do patients themselves demonstrate different manifestations, but their clinical outcomes can be drastically different. A typical example is COVID-19 patients with comorbidities such as type-II diabetes are usually more severe.

However, most commonly, the study of medicine focuses usually on a single phenotype of their interest while ignoring comorbidities. Within the big data era, more and more data from the EHR is now available to unleash the potential of studying how diseases form clusters and influence the clinical outcomes of each other.

Until the most recent decade, researchers started to explore the heterogeneity of previously defined phenotypes and uncovered novel distinct patterns within the previously well-defined phenotypes. These works demonstrated the power of unsupervised machine learning in pattern recognition and opened a new and interdisciplinary field -- the study of comorbidity and its influences on clinical outcomes.

In this work, we will utilize the embedded patient vectors to perform unsupervised clusters and reveal heterogeneity within a single phenotype. We will characterize each cluster group and identify their key attributions to represent their unique clinical features and explore related clinical outcomes.

6.2 Related Work

There are not many studies that focus on illustrating the heterogeneity of diseases, as it usually requires a large amount of data and noticeable patterns. Most recently, with the advancement of unsupervised machine learning and big data, some studies have explored the potential of identifying heterogeneity cluster groups within the diseases of their interest.

A group in early 2020 used an unsupervised learning method, Latent Dirichlet Allocation (LDA), and a generative probabilistic model called the Poisson Dirichlet Model (PDM) to study the cluster of osteoporosis, dementia, and COPD in the EHR. They identified survival differences and a few demographics among these cluster groups. This group focused on three complex diseases and used an unsupervised approach to demonstrate phenotype discrepancies [14].

Just a few months later, the group designed Deep Patient, applied another unsupervised approach, and studied eight different complex diseases, identifying cluster groups within pre-defined phenotypes. Their work utilized a convolutional neural network plus an autoencoder to perform a patient embedding process before clustering. However, their work had little discussion on the clinical outcomes or other characteristics besides some level of comorbidities within the clusters [12].

In sum, the field of identifying heterogeneity within a defined phenotype is still new. Most studies still focus on a single phenotype without further consideration of the internal discrepancies of patients. However, the above studies have explored and identified certain variations of individuals within a certain phenotype. As individuals do encounter a certain amount of diseases and gain a wide variety of phenotypes, we reason these internal differences are crucial and can provide value for personalized medicines.

6.3 Methods

6.3.1 Clustering patients using Gaussian Mixture Model (GMM)

We use GMM to perform the cluster based on the 50 features of the embedded patient vector. GMM is a mixture model that uses a combination of multiple Gaussian densities to interpret the given distribution. GMM is parametric, which can be seen as a weighted combination of Gaussian distribution. Similarly, K-means clustering is another clustering algorithm, but non-parametric, without any specific assumptions about the data distribution. Here, we adopt GMM for the clustering task because the output layer of embedding uses sigmoid functions. Thus, a cluster of patients is a mixture of sigmoid signals, which can be approximated by a mixture of Gaussians. The parameter estimations for a GMM are computed iteratively using Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation.

6.3.2 Defining the reasonable cluster number using Bayesian Information Criteria (BIC)

In theory, given a finite number of data points, there would be numerous numbers of models to fit the data. To select the ideal model (i.e., the best model) that fits the data, statisticians have developed many model evaluation or comparison metrics. Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are two commonly used metrics to compare model fit. AIC, developed by Japanese statistician, Hirotsugu Akaike, uses the likelihood function of a model L and the numbers of estimated parameter k to compute the AIC.

While BIC, developed based on AIC by Gideon E. Schwarz, is very similar to AIC but added a greater penalty term on increased sample size.

Though there is no formal statement on which is better to use, we choose BIC in our work, as we hypothesize that adding a penalty term on sample size should fit better as we include more data points.

We use BIC to select the ideal number of clusters. For a possible number of cluster groups (here, we define from a minimum of 2 clusters to a maximum number of 10 clusters), we experiment from the minimum to maximum numbers of clusters and record their BIC score. Usually, we reasonably select the lowest BIC score as the ideal number of clusters. However, sometimes if the BIC score of cluster number 5 is barely the same as cluster number 6, we will simply avoid over-diving our samples by choosing the smaller number of clusters, in this case, cluster 5. We will demonstrate the use of BIC in the result session.

6.3.3 Compute phenotype enrichment using logistic regression model

We performed comorbidity analysis within a single phenotype to reflect the heterogeneity. GMM is used to first group samples into clusters. We chose the number of clusters based on the Bayesian information criteria (BIC).

To characterize the distinct feature within each cluster group, we performed logistic regression for each comorbidity (defined by PheCode), including the cluster (using one versus the rest), age of onset, sites, gender, ethnicity, and race as covariates to predict the comorbidity.

Theoretically, this step could be placed by simply computing the concurrences between phenotypes using Fisher's exact test. However, the dataset we used is from the eMERGE network, which collected data from more than 10 clinical sites. There is potential ascertainment bias and differences within each site. Thus, we implemented the logistic regression model and accounted for various covariates to test the enriched phenotypes within each cluster.

We retrieved phenotypes that are significantly associated with each cluster and used these phenotypes as cluster-associated comorbidities to characterize each cluster group. Statistically, we adjusted for multiple tests (n = 1,855 comorbidities) using Benjamini-Hochberg adjusted p-values < 2e-5 as the significance level.

6.3.4 Principal component analysis of longitudinal embeddings

To understand and decompose the variation of the longitudinal progression in specific phenotypes, we first used principal component analysis (PCA) to linearly reduce the dimension of the longitudinal vector.

PCA is a powerful statistical technique widely used in the fields of modern data analysis. The primary goal of PCA is dimensionality reduction. PCA transforms high-dimensional data into a lower-dimensional representation, using a linear combination of existing features. After PCA transformation, the original data is represented in predefined numbers of principal component (PC) space, for example, PC1 and PC2. Due to the orthogonality property of the transformation, PCs are naturally orthogonal to each other, with PC1 explaining most of the variances.

Mathematically, PCA can be conducted using singular value decomposition (SVD), a matrix operation method. SVD decomposes a given metric into three metrics, denoted as X=UΣV, with X as the original metric. In the decomposed matrix, V is the principal component (also known as eigenvectors).

6.3.4 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) is done first by fitting a linear regression model including sex, age of onset, site, and race as covariates, using clusters to predict the first three PCs, respectively. Then, according to the linear mode's results, ANOVA performed F-tests to evaluate the null hypothesis that all groups have the same mean by comparing the mean square between (MSB) and mean square within groups (MSW). Where:

$$\frac{MSB}{MSW}$$

And

$$MSB = \frac{SSB}{k-1}$$

$$MSW = \frac{SSW}{n-k}$$

Where SSB indicates sum square between and SSW indicates sum square within groups. Below is the formula of calculating SSB and SSW given *k* groups and *n* samples in total (with n_i indicates sample size in group *i*).

$$SSW = \sum_{j=1}^{n} \sum_{i=1}^{k} (\bar{y_i} - y_j)^2$$

$$SSB = \sum_{i}^{k} n_i (y - \bar{y_i})^2$$

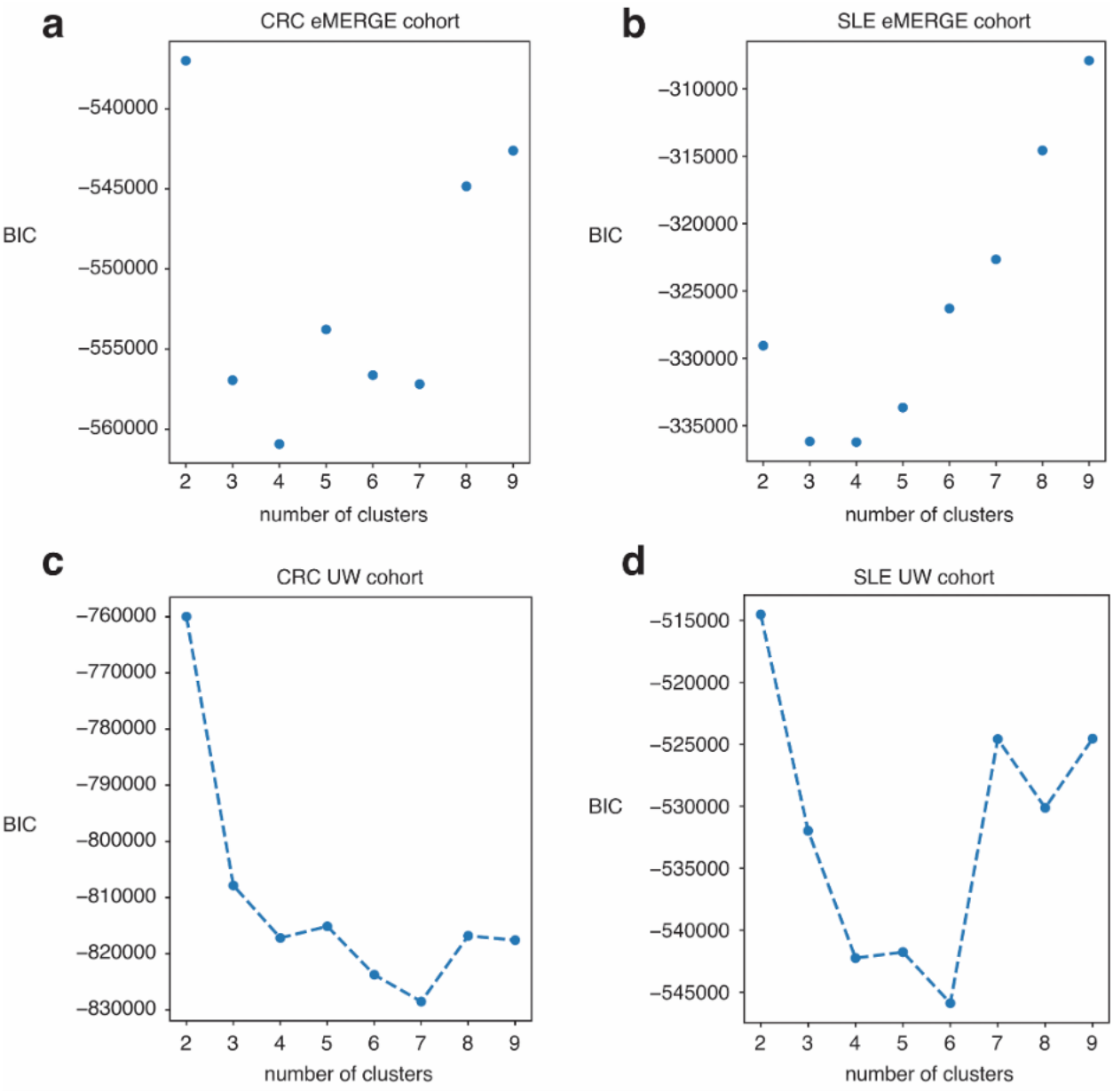
6.4 Results

In the result session, we will present the study on two phenotypes, including systemic lupus erythematosus (SLE) and colorectal cancer (CRC). This session will provide a detailed characterization of these two phenotypes and discuss clinical features and potential impact.

6.4.1 Systemic lupus erythematosus heterogeneity and comorbidity differences

Systemic lupus erythematosus (SLE) is a chronic autoimmune disease that can affect multiple organs and tissues in the body (that's why it is called systemic), including the skin, joints, kidneys, brain, and other internal organs. Current medical knowledge can fully dissect the exact cause of SLE, but it is believed to involve a combination of genetic, environmental, and hormonal factors. The severity of SLE can vary across patients. There is no cure for SLE now. Treatment usually targets the symptoms, trying to reduce the inflammation and thus reduce the self-tissue damage.

Figure 6.1 Bayesian Information Criteria (BIC) curve for model selection. (a, b) for eMERGE cohort and (c, d) for the UW cohort

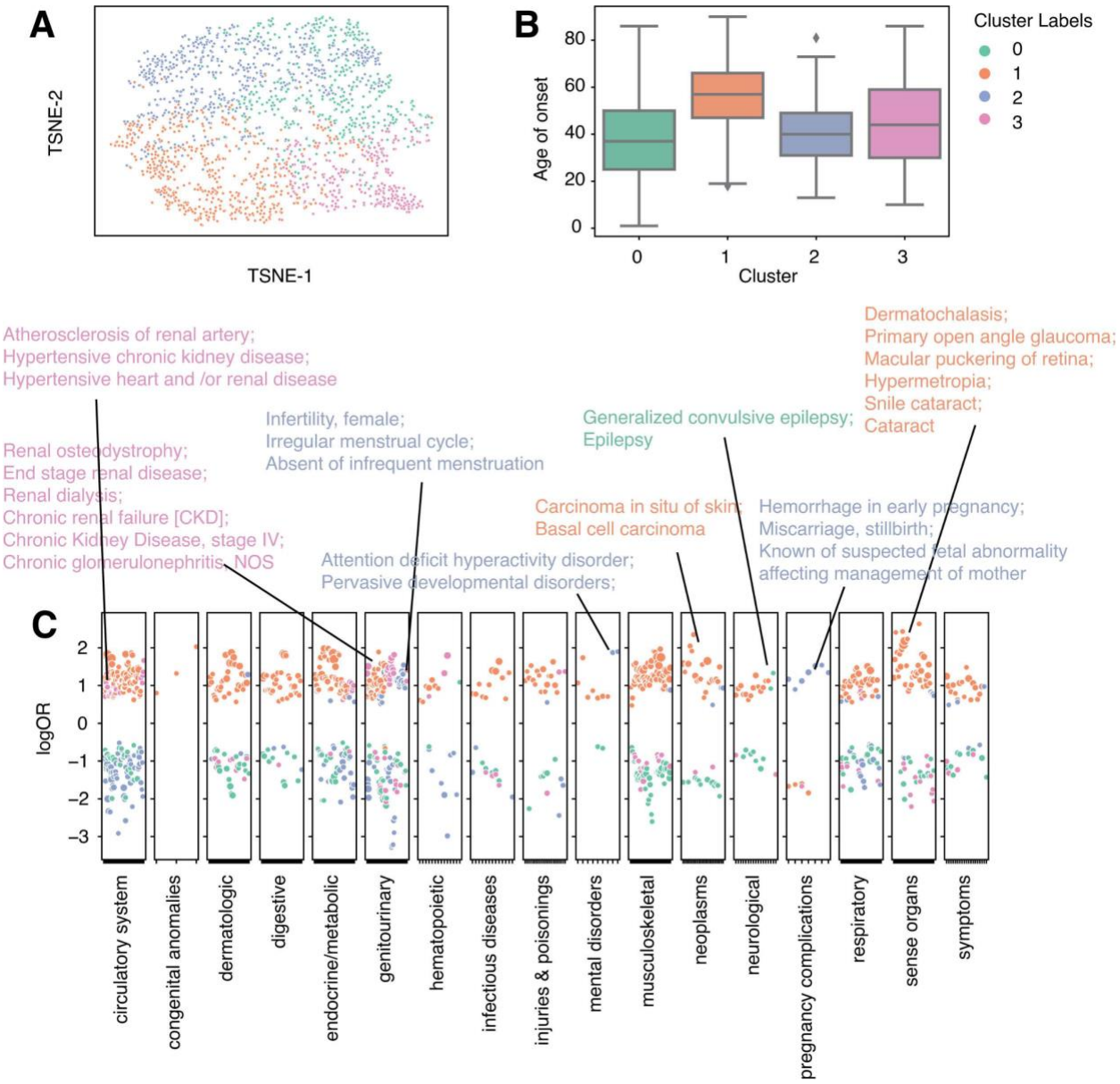


In this work, we performed GMM clustering on systemic lupus erythematosus (SLE) patients (n = 1806). We identified 4 clusters according to BIC and observed a wide range of disease patterns within these clusters (Figure 6.1). Cluster 0 has the lowest onset age (median onset age = 37) among all other groups and is associated with epilepsy. Cluster 1 has the highest median onset age (median onset age = 57) and is joint with skin cancer and eye diseases, such as glaucoma, cataracts, dermatochalasis, etc. Evidence shows

that SLE patients can develop cataracts and many other eye diseases [79,80], and some might be due to medications [80]. Cluster 2 has a median onset age similar to cluster 0 (median onset age = 40) and is enriched in pregnancy complications, including hemorrhage in early pregnancy, miscarriage, stillbirth, etc. Cluster 2 also has signs of infertility, irregular menstrual cycle, and developmental disorder (Figure 6.2). Though still unclear, numerous studies have tried to dissect the relationship between lupus and pregnancy complications and identified hormone-level abnormalities [81,82]. Cluster 3 has a median onset age of 44 and is associated with renal diseases, including renal osteodystrophy, end-stage renal disease, chronic kidney diseases, etc. SLE is known to be a systemic disease, and cluster 3 reflects its systemic involvement in kidney disorder. When aggravated, it can lead to kidney failure. Using CRC and SLE as two case studies, we show that patient vectors can reveal distinct comorbidity patterns. Even within a single phenotype, there are diverse patterns of comorbidities. Thus, further evaluation and personalized care plan is required to improve healthcare.

Figure 6.2 Clustering analysis identified subgroups with distinct comorbidity patterns in Systemic lupus erythematosus patients (n = 1806) from the eMERGE cohort.

A. TSNE plot of patient vectors colored by cluster groups defined using Gaussian Mixture Model (GMM) with optimal Bayesian Information Criteria (BIC).
B. Box plot showing distribution of age of onset for individual SLE cluster groups.
C. Comorbidity pattern enrichment plot grouped by disease classes (in x-axis) within each cluster group (represented by color). The y-axis indicates the log odds ratio of the comorbidity enrichment. Only statistical significant results are shown ($p < 2e-5$) after Bonferroni correction. Colored texts are used to highlight the top results within each cluster group.



6.4.2 Colorectal cancer heterogeneity and comorbidity differences

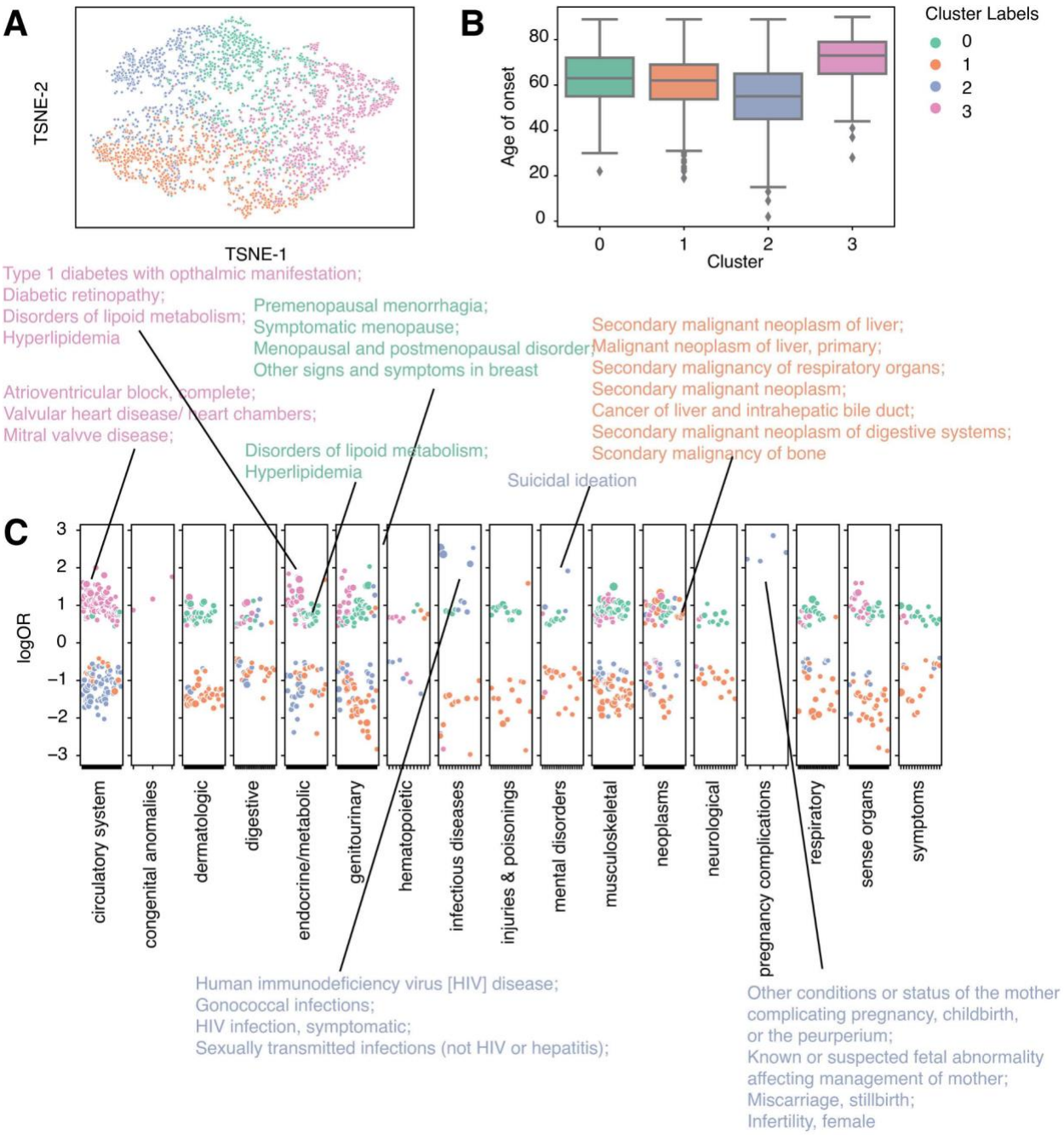
Comorbidity, the simultaneous presence of two or more diseases or medical conditions, has a profound impact on an individual's care plan, quality of life, and mortality [32]. Using the summed-up embedding to represent individual patients, we show that within a well-defined single phenotype, the comorbidity status formed different clusters, showing heterogeneity in comorbidity patterns (Figure 6.3). Using colorectal cancer (CRC) patients (identified using phecode 153) as an example (n = 2837), we identified 4 clusters

according to the Bayesian Information Criterion (BIC) using Gaussian Mixture Models (GMM) (Figure 6.1). To characterize the comorbidity patterns within each cluster group, we fitted logistic regression models using one cluster group versus the rest strategy adjusted for age of onset, race, ethnicity, and sites (see Methods). Cluster 2 (median onset age = 51) has the youngest age of onset and is strongly associated with HIV infection (PheCode = 071) and a few pregnancy complications, representing a subgroup of female patients with immunodeficiency phenotypes. This association between HIV and CRC is not new and is more prevalent in women in a pooled result from 3 studies [83–86]. Cluster 1 (median onset age = 62) is associated with secondary malignant neoplasm, which reflects cancer pleiotropy and late-stage cancer patients with metastasis [87,88]. Cluster 0 (median onset age = 60) is enriched in genitourinary and endocrine diseases, including disorders of lipoid metabolism, menopause issues, and menorrhagia issues. Though endocrine and metabolic disease might be risk factors for CRC, and vice versa, a study has also shown a greater risk of CRC patients developing endocrine and metabolic diseases [89,90]. Cluster 3 (median onset age = 72) group is the latest onset group, which has a strong pattern of the circulatory system and endocrine diseases, including atrioventricular block, valve heart disease, mitral valve diseases, diabetics, and hyperlipidemia. This cluster group aligns with existing findings that CRC patients have an increased risk of developing cardiovascular disease and heart failure [91,92].

Figure 6.3 Clustering analysis identified subgroups with distinct comorbidity patterns in colorectal cancer patients (n = 2837) from the eMERGE cohort.

A. TSNE plot of patient vectors colored by cluster groups defined using Gaussian Mixture Model (GMM) with optimal Bayesian Information Criteria (BIC).
B. Box plot showing distribution of age of onset for individual CRC cluster groups.
C. Comorbidity pattern enrichment plot grouped by disease classes (in x-axis) within each cluster group (represented by color). The y-axis indicates the log odds ratio of the comorbidity enrichment.

Only statistical significant results are shown ($p < 2e\text{-}5$) after Bonferroni correction. Colored texts are used to highlight the top results within each cluster group.



6.4.2 Validation using EHR data from University of Washington

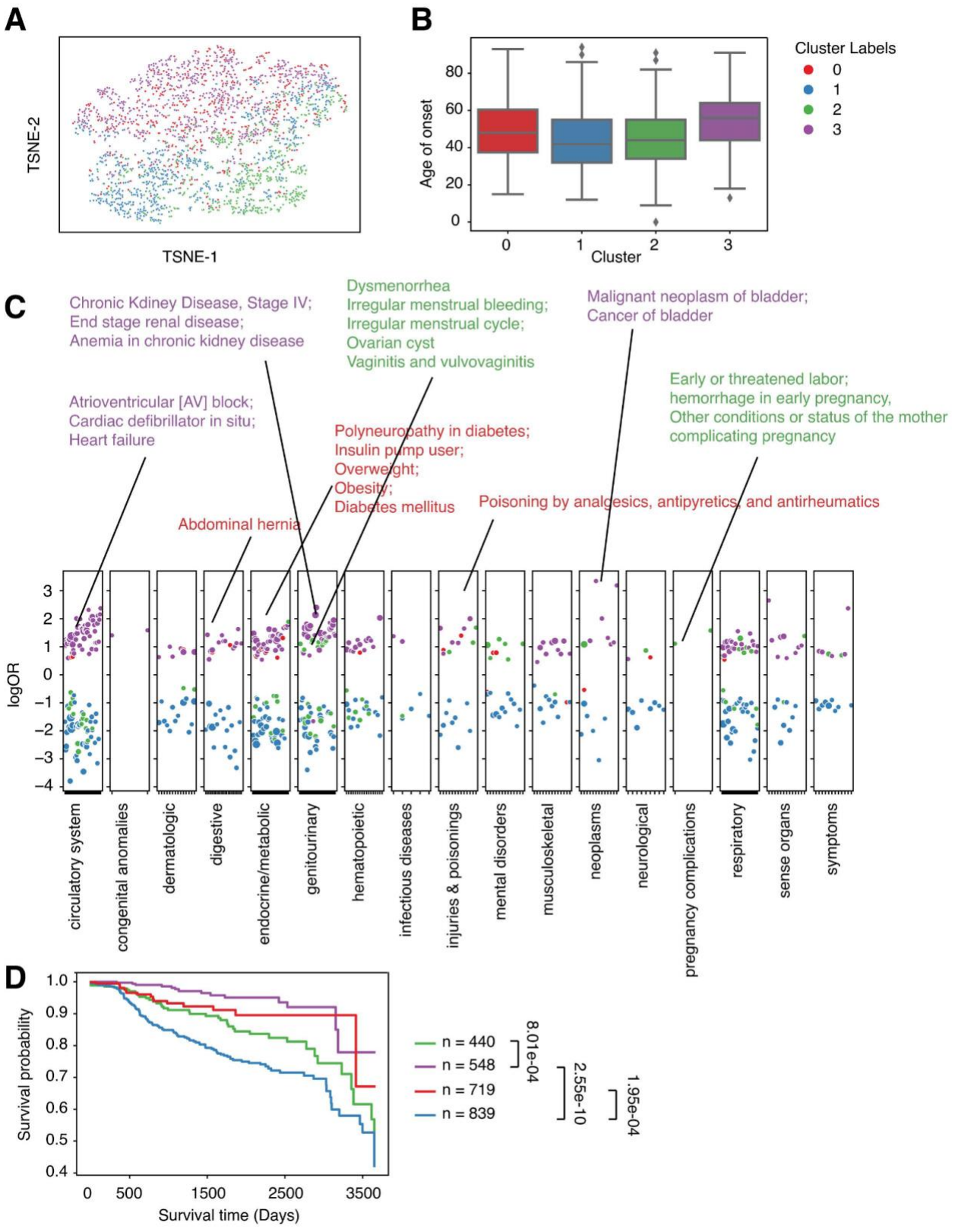
In this session, we report the heterogeneity analysis on the local UW EHR data, serving as an external validation. The goal of this session is to see to what extent we can reproduce the identified cluster groups. The reproducible results of the cluster group on another dataset (UW local EHR data) can demonstrate the stability of these cluster groups and support these clinically meaningful features for further studies.

We again utilized SLE and CRC as examples, aiming to reproduce the similar results we observed in the eMERGE dataset. In the UW validation cohort, we identified 4 clusters of SLE enriched for different comorbidities (Figure 6.4). Cluster 2 (median age of onset = 44) enriched with pregnancy complications, genitourinary, and a few mental disorders, has a relatively early onset age and is highly similar to what we identified in the eMERGE cohort. Cluster 0 (median age of onset = 48) is associated with endocrine/metabolic (such as diabetes, overweight, and obesity). Cluster 3 (median age of onset = 56) is associated with the circulatory system (Atrioventricular block, Cardiac defibrillator in situ, Heart failure, etc.), genitourinary (Chronic Kidney diseases, End stage renal disease, Anemia in chronic kidney disease, etc.), and a few neoplasms (Malignant neoplasm of bladder, Cancer of bladder). Cluster 1 (median age of onset = 42) does not have a unique pattern enrichment of comorbidities. In short, we identified 4 clusters of SLE in the UW cohort, with three having distinct disease patterns. Two cluster groups have identical properties to what we have found in the eMERGE dataset, including pregnancy-complication-associated lupus (Cluster 2) and Renal-associated lupus (Cluster 3). Additionally, with available survival data in the UW cohort, we compared the 10-year overall survival among different cluster groups (Figure 6.4D). Among these comorbidity groups, cluster 1, which showed no comorbidity enrichment, has the lowest survival rate in 10 years. We reason that the low survival rate might partially be explained by a higher proportion of males in cluster 1 compared to other clusters (odds = 2.54, p = 1.42e-14). Though not widely reported, we observed a lower life expectancy in male SLE patients vs. females. Then, we noticed that cluster 2 also showed a relatively lower survival rate than Cluster 3 (FDR = 8.01e-4). This result might imply that

pregnancy-complication-associated lupus patients might need more follow-up and on-time treatment to improve their health outcomes and life expectancy. One national study also found that SLE women have 20-fold higher maternal death [81]. Clusters 0 (Diabetes-associated SLE) and Cluster 3 (Renal-associated SLE) showed a better survival status, which also has a relatively late onset age, representing late-stage SLE when patients age since SLE usually involves multiple organ-level dysfunctions.

Figure 6.4 Clustering analysis identified subgroups with distinct comorbidity patterns in SLE patients (n = 2546) from the UW validation cohort.

A. TSNE plot of patient vectors colored by cluster groups defined using Gaussian Mixture Model (GMM) with optimal Bayesian Information Criteria (BIC).
B. Box plot showing distribution of age of onset for individual CRC cluster groups.
C. Comorbidity pattern enrichment plot grouped by disease classes (in x-axis) within each cluster group (represented by color). The y-axis indicates the log odds ratio of the comorbidity enrichment. Only statistical significant results are shown ($p < 2e-5$) after Bonferroni correction. Colored texts are used to highlight the top results within each cluster group.

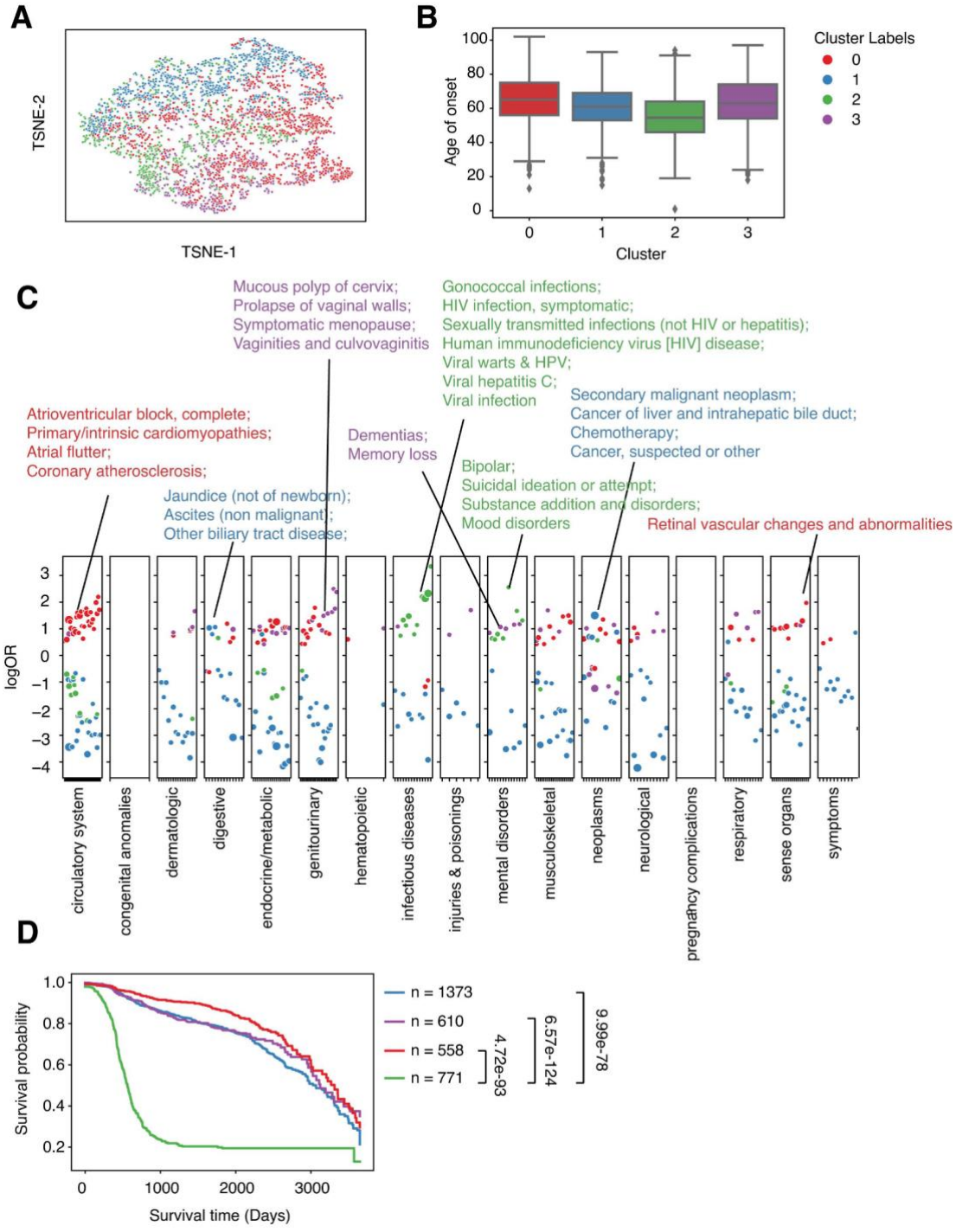

A
TSNE-2
TSNE-1
B
Age of onset
Cluster
Cluster Labels
0
1
2
3
C
Chronic Kdiney Disease, Stage IV;
End stage renal disease;
Anemia in chronic kidney disease
Dysmenorrhea
Irregular menstrual bleeding;
Irregular menstrual cycle;
Ovarian cyst
Vaginitis and vulvovaginitis
Malignant neoplasm of bladder;
Cancer of bladder
Atrioventricular [AV] block;
Cardiac defibrillator in situ;
Heart failure
Polyneuropathy in diabetes;
Insulin pump user;
Overweight;
Obesity;
Diabetes mellitus
Early or threatened labor;
hemorrhage in early pregnancy,
Other conditions or status of the mother
complicating pregnancy
Abdominal hernia
Poisoning by analgesics, antipyretics, and antirheumatics
logOR
circulatory system
congenital anomalies
dermatologic
digestive
endocrine/metabolic
genitourinary
hematopoietic
infectious diseases
injuries & poisonings
mental disorders
musculoskeletal
neoplasms
neurological
pregnancy complications
respiratory
sense organs
symptoms
D
Survival probability
Survival time (Days)
n = 440
n = 548
n = 719
n = 839
8.01e-04
2.55e-10
1.95e-04

We also discovered 4 clusters of CRC in the UW cohort (Figure 6.5). Likewise, one cluster (cluster 2, the median age of onset = 54.5) that has a relatively younger age of onset is enriched in infectious diseases (HIV infection, Viral hepatitis C, etc.) and a few mental disorders (Bipolar, Suicidal ideation or attempt, Mood disorder, etc.), identical to cluster 2 from the eMERGE cohort. Cluster 0 (median age of onset = 65) from the UW cohort is similar to cluster 3 from the eMERGE cohort, as both showed phenotype enrichment in circulatory systems (Atrioventricular block) and endocrine/metabolic diseases. Moreover, cluster 1 (median age of onset = 61) from the UW and eMERGE cohorts are both enriched in secondary malignant neoplasm, specifically, cancer of the liver and intrahepatic bile duct. Together, these results provided compelling evidence that our findings from the training cohort (the eMERGE cohort) are stable and can be validated using the UW cohort externally. Again, analyzing the 10-year overall survival difference (Figure 6.5D), we found cluster 2 enriched in infectious diseases and mental disorders showed a significantly lower overall survival probability than the other 3 clusters. Searching through existing literature, though there are a few discussions about HIV and colorectal cancer risk, only one report used a meta-analysis method investigating the mortality rate of CRC patients with HIV with non-significant results, partially due to inadequate cases (n = 194) [83]. One report also found HIV-infected cancer patients with elevated mortality rates [93]. Our findings provide extra evidence supporting these research works and demonstrate the potential of uncovering new patterns in clinical outcomes among patients. Also, given that most research currently lacks the consideration of comorbidity patterns in outcome prediction and personalized medicine, our data shows that comorbidity patterns can provide crucial information in patient care and life expectancy.

Figure 6.5 Clustering analysis identified subgroups with distinct comorbidity patterns in CRC patients (n = 3673) from the UW validation cohort.

A. TSNE plot of patient vectors colored by cluster groups defined using Gaussian Mixture Model (GMM) with optimal Bayesian Information Criteria (BIC).
B. Box plot showing distribution of age of onset for individual CRC cluster groups.

C. Comorbidity pattern enrichment plot grouped by disease classes (in x-axis) within each cluster group (represented by color). The y-axis indicates the log odds ratio of the comorbidity enrichment. Only statistical significant results are shown ($p < 2e\text{-}5$) after Bonferroni correction. Colored texts are used to highlight the top results within each cluster group.



6.4.3 Progressional analysis of CRC patients in the eMERGE cohort

Another question we thought to ask is do these patients form clusters at the moment we saw them. Or do those cluster groups reveal different trajectories that lead them to what they are now? To understand this, we further evaluated the potential of using longitudinal embedding vectors to study the progression of diseases. We used CRC (PheCode 153) as an example, using all available patients with 10-year longitudinal vectors after a diagnosis of CRC (n = 110). We first performed Principal Components Analysis (PCA) on the longitudinal vectors, trying to decompose the changes in disease progression and analyze the variance. We included the first three PCs, which composed 51% of the explained variances (Figure 6.6). Using analysis of variance (ANOVA), including the age of onset, race, gender, site, and clusters as covariates, we found that PC1 and PC2 are explained majorly by the cluster groups we identified using GMM (see session), then the age of onset, gender, sites, and races (Figure 6.6A, B). And PC3 is explained mainly by sites (Figure 6.6B). This result suggests that the variation in disease progression longitudinally is also captured by the clusters, indicating that individual cluster groups also have a different disease progression track, meaning that the cluster groups we identified can project the disease trajectories. To understand the differences in progression, we then analyzed the emerging phenotypes following the onset of the disease, revealing substantial differences among cluster groups (refer to Figure 6.6C). Besides the consistent occurrences of "Malaise and fatigue" and "Other anemias" across all four clusters, a few phenotypes were also present in three out of four clusters. These included "Essential hypertension," "Gastrointestinal hemorrhage," "Other symptoms of the respiratory system," "Benign neoplasm of the colon," and "Abdominal pain." The remaining emerging phenotypes displayed radical variations across all four clusters, indicating distinct progression trajectories and varied comorbidity patterns.

Given the substantial disparities in the progression of the four clusters, we investigated their disease patterns before the onset of the CRC. Our investigation revealed their initial divergence, as illustrated in Figure 6.7. However, these disparities are not particularly prevalent; the maximum frequency across all four clusters is

merely 21%. This frequency can subsequently escalate to as high as 55% after the onset of the disease. Together, these findings imply that disease progressions exhibit a high degree of heterogeneity, yet they still manifest discernible patterns in terms of comorbidities. This observation indicates varying disease risks and underscores the potential for personalized medicine approaches.

Figure 6.6. Longitudinal analysis revealed progressional differences within each cluster group in CRC.

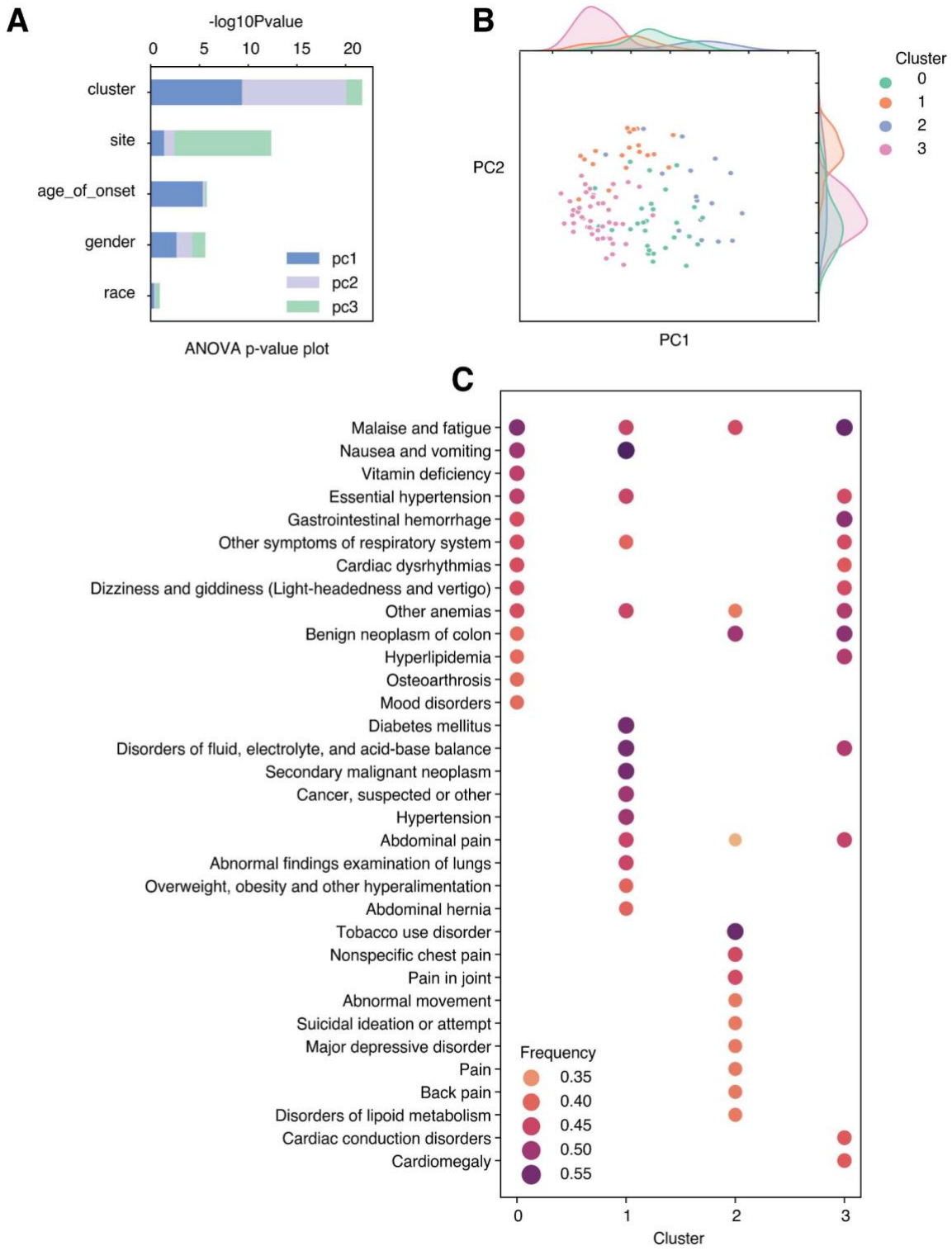
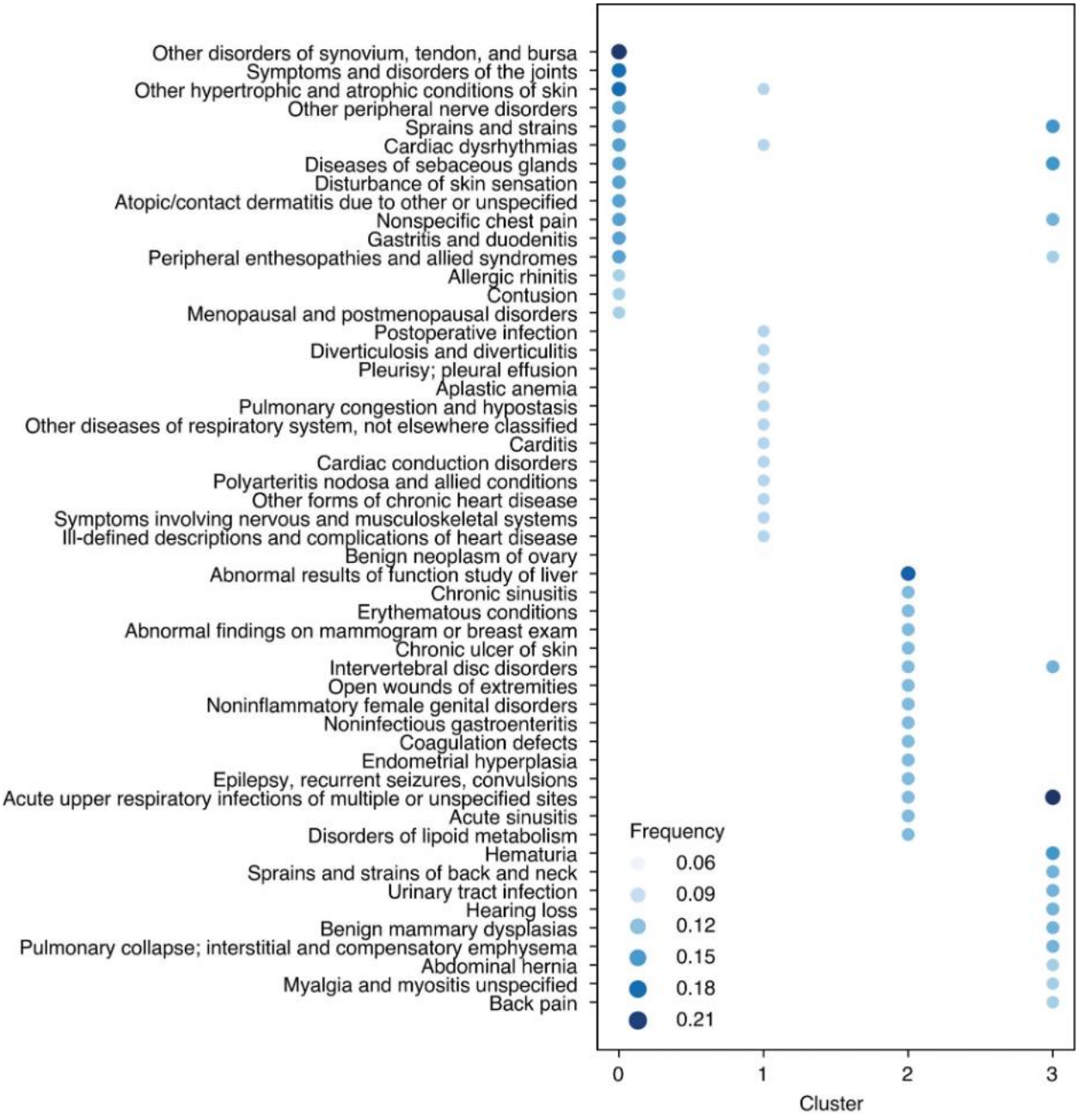


Figure 6.7 Comorbidity differences and prevalence before the onset of CRC across each cluster.



6.4.4 Online deployment of the EHR phenotype clustering

As we demonstrated the validity of these analyses using two phenotypes as examples, with external validation from UW EHR data, we deployed the clustering analysis of many other phenotypes online, for

others to investigate further topics of interest. The web-deployment is currently hosted on https://ehrcluster.web.app/.

6.5 Conclusion

Here, with a comprehensive exploration in Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors), we utilized the patient embedding with clustering analysis to study the heterogeneity within CRC and SLE, answering the question of studying disease patterns using EHR data. These analyses revealed novel disease heterogeneity patterns with meaningful clinical interpretations and literature support. Though some of these novel patterns are not commonly observed and the associated overall mortality is not validated to be causal, these results serve as pioneer studies for further interrogation of the complex relationships, which can finally be utilized in personalized medicine. At this point, we have explored the secondary use of EHR using unsupervised learning and demonstrated the use of the EHR data to identify novel disease patterns, which is the first and major part of the thesis goal.

In this chapter, we explored the use of patient embedding to analyze patient progression trajectory and comorbidity heterogeneities. There is a long hold of doubt as to why individuals respond differently to treatment and exhibit discrepancies in symptoms and comorbidities, even though they are diagnosed with the same disease. Here, we utilized this unsupervised clustering method to analyze the discrepancies and revealed several differences in patients with the same diseases, including age of onset, comorbidities, progressions, and mortality.

In disease comorbidity analysis, we applied cluster algorithms and detected distinct comorbidity patterns within CRC and SLE. Additionally, we reproduced similar cluster results and revealed distinctions in overall survival in the UW cohort. In the progression analysis, we show that each cluster is associated with

distinct phenotype gain, which suggests that the cluster groups are progressively different, meaning that we are not capturing a static moment but a continuous variation.

Together, these results suggest that partitioning patients into different subgroups can help identify disparities in disease progression and critical health outcomes using EHR data. Though more efforts are needed to understand the complex comorbidity relationship, implementing an intelligent clinical decision support system to facilitate personalized medicine is feasible, leveraging the abundant patient data within the EHR. For instance, when the system notices that an HIV patient has recently been diagnosed with CRC, which is an extremely high risk, an early warning can inform the severity of the co-occurrence of these two phenotypes. This system can help identify urgency and achieve early treatment to increase life expectancy. Besides actions in clinical application, identifying comorbidity patterns within a phenotype might indicate pathological or etiological similarities between co-occurring phenotypes. These comorbidity patterns can facilitate fundamental scientific advances in identifying molecular signatures of diseases.

To provide a more comprehensive atlas of this analysis, we deployed an online visualization of the cluster analysis for almost 1000 different phenotypes online, allowing individual researchers to explore and study based on their expertise. We believe this can further drive our understanding of disease heterogeneity to the next level.

In this chapter, we addressed the objective of Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors) and finally demonstrated the use of patient embeddings to uncover novel disease patterns. In the following chapter, we will focus on Aim 3 (to perform a genome-wide association study (GWAS) for depression phenotype using rule-based phenotyping algorithm derived from the EHR), studying disease etiology using EHR data.

Chapter 7: EHR derived Depression phenotyping algorithm and GWAS

7.1 Overview

In all previous chapters, we explored the secondary use of EHR from the unsupervised method and assessed the potential of utilizing the EHR data to identify novel disease patterns. This chapter will illustrate a supervised approach (rule-based method) to study disease etiology. The previous chapters outlined a pioneering approach using patient embeddings to identify novel disease patterns, fully addressing the objectives of Aim 1 (to develop patient representation learning in EHR data using an unsupervised machine learning approach) and Aim 2 (to identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors). To explore disease etiology, the EHR offers a rich source of patient data. Commonly, disease etiology encompasses behavioral, genetic, and environmental factors. Extensive evidence has already established numerous genetic associations with various diseases. Aligned with Aim 3 (to perform a genome-wide association study (GWAS) for depression phenotype using rule-based phenotyping algorithm derived from the EHR), in this chapter, we focus on constructing a refined depression cohort using an EHR-derived phenotyping algorithm to investigate genetic risk factors. In this work, we explored the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology

As we mentioned before, for the time being, machine learning algorithms don't ensure a lead when compared to traditional rule-based phenotyping algorithms. Machine learning algorithms provide the benefit of high-throughput processes, as it is intrinsically automatic. However, for complex diseases, machine learning can fall off, as the discovery and the definition of such complex diseases are still evolving

with the gain of new medical knowledge. For instance, psychiatric diseases are becoming more complicated as they involve sophisticated interactions among environment, genetics, and lifestyles.

Depression is a pervasive mental health disorder. Depression usually has symptoms of feeling sad, loss of interest, and other cognitive and even physical symptoms. Depression affects millions of individuals worldwide and can cause significant personal suffering, impaired functioning, and socioeconomic burden [94]. Current research suggests that depression is a complex and multifaceted disease. The etiology of depression includes an interaction of genetic, environmental, and psychological factors [95][96]. While environmental stressors and psychosocial factors have been extensively studied, genetic factors have emerged as key contributors to the susceptibility and development of depression [97].

Genome-wide association study (GWAS) is one of the approaches to uncovering the genetic basis of complex psychiatric disorders, including depression [98]. GWAS enables a comprehensive exploration of the entire genome, scanning thousands of genetic markers in large cohorts of individuals with depression and controls. These studies have identified a large number of genomic regions associated with depression susceptibility, implicating various biological pathways, including neurodevelopment, neurotransmission, and immune response [97][99]. Nonetheless, the effect sizes of the identified variants have been modest, indicating that depression is likely influenced by a multitude of genetic variants, each conferring a small effect size.

Understanding the genetic basis of depression is of great importance and clinically meaningful. First, genetics can provide crucial insights into the underlying biological mechanisms, potentially leading to the development of more targeted and effective therapeutic interventions. Second, genetic studies can help identify individuals at higher risk of developing depression, facilitating early detection and intervention. Lastly, the genetic exploration of depression can contribute to reducing the stigma associated with the disorder by emphasizing its biological nature.

For decades, studies have attempted to link depression with immune dysfunctions [100]. Until most recently, researchers recognized the interplay between depression, chronic stress, and weakened immune responses [101]. These findings start to show that the impact of depression does not stay merely within the psychiatric domain but is more systemic. The Human Leukocyte Antigen (*HLA*) region, also known as the Major Histocompatibility Complex (MHC), is a crucial genetic complex located on the short arm of chromosome 6 in humans. This region plays a fundamental role in the immune system by encoding a set of cell surface proteins that are essential for the recognition of self and non-self antigens. The *HLA* system is highly polymorphic, meaning that there is a vast diversity of *HLA* alleles within the human population.

The primary function of the *HLA* molecules is to present antigens to T cells, a critical step in the immune response. Antigens are molecules that the immune system recognizes as foreign, such as those derived from pathogens like bacteria or viruses. *HLA* molecules, which are expressed on the surface of nearly all nucleated cells in the body, bind to these antigens and present them to T cells for inspection. With the advance of sequencing technology and knowledge about the human genome, the study of the *HLA* region has identified many associations between *HLA* alleles and immune-related diseases.

Most of the GWAS studies have used diagnosis codes or expertise-based manual effort to build the cohort, due to a lack of a unified phenotyping method [102][103]. Thus, it is crucial to build a standard phenotyping algorithm to identify depression in the clinical environment. In this study, we aim to contribute to the growing body of knowledge on the genetic basis of depression by conducting a GWAS on an EHR-derived depression cohort using data from the eMERGE Network [7,104]. By employing state-of-the-art genetic and functional analysis tools, we identified and annotated novel genetic variants in the MHC-II and IGHV regions associated with depression susceptibility. We will focus on the molecular mechanisms of depression and link it with genetics.

7.2 Related Work

Depression is characterized as a psychiatric disease. However, some clinical evidence linked immune response and chronic inflammation phenotypes to depression. Back in 1993, a meta-analytic review unraveled the relationship between clinical blood markers and depression, showing a decreased proliferative response of lymphocytes to mitogens, natural killer cell activity, and several white blood cell populations [100]. A recent study using polygenic risk score (PRS) of depression also identified a strong association between white blood cell counts and the PRS [105]. It is becoming more accepted that depression as a psychiatric disease can affect the immune system. A few genetic studies identified genetic variations associated with depression risks [97,102]. We will evaluate the depression genetic risk linking, specifically within the *HLA* region, using the eMERGE data.

The study of diseases relies heavily on expertise within the domain to provide insights and usually requires a great effort to collect, characterize, and define a curated cohort. The eMERGE consortium designed and implemented an EHR-based depression phenotyping algorithm, using various diagnoses, procedure codes, medications, and a series of strict rules to facilitate the study of depression [6,106]. To our knowledge, this is the first EHR-based depression phenotyping algorithm. Enabling a systematic and clean EHR-based phenotype can enhance community awareness and improve the reproducibility of genetic findings by providing a unified phenotyping algorithm for complicated phenotypes. As a psychiatric disorder with various levels of symptoms, depression is commonly underdiagnosed [107]. In some severe cases, depression can be lethal [108]. Contemporary studies revealed that depression involves many aspects, including the immune system, neurological disorder, and genetics. In this work, we will first use the depression phenotype to conduct a Genome-Wide Association Study (GWAS) to 1). Replicate existing genetic risks of depression to evaluate the validity of the EHR-based depression algorithm. And 2). Identify potential new genetic risks to explore the complicated molecular mechanisms of depression.

7.3 Methods

7.3.1 Data collection and cleaning

We used eMERGE phase I-III imputed genotype data and the curated EHR-based depression phenotypes to perform the GWAS analysis. The combined ancestry data contains around 50k samples of European, African, and Asian ancestry participants. The phenotype data is collected from PheKB, available at https://phekb.org/node/1095/implementations-table. The GWAS analysis combined all three depression phenotypes (non-major depression, major depression, and major depression with psychosis) as cases, due to insufficient sample sizes when breaking down.

7.3.2 Genotyping and Imputation

Genotypes for all participant samples from eMERGE-I, eMERGE-II, and eMERGE-III were imputed using the Michigan Imputation Server, using the Haplotype Reference Consortium reference panels [109] [110] (HRC1.1). Most samples were genotyped with the Human 660 Quad. Other genotyping platforms included the CytoSNP-850K BeadChip, the OmniExpress chip, the Affymetrix 6.0 array, and the Illumina MEGA.

7.3.3 Genetically determined ancestry (GDA)

The genetic determined ancestry is inference from the PCA results. We performed PCA analysis on all participant samples from eMERGE-I, eMERGE-II and eMERGE-IIIusing the PLINK 2.0 software [111]. Variants with >= 0.05 MAF, missingness of <= 0.1 and LD-pruned R^2 threshold of 0.7 were included. The genetically determined ancestry (GDA) was defined by K-means clustering of Principal Component (PC) 1 and PC2 and three groups (corresponding to African ancestry, Asian ancestry and European ancestry) were identified.

7.3.4 GWAS analysis

The GWAS analysis included the following covariates: 1. Decade of birth (rounded as integer); 2. Median BMI; 3. Sex; 4. Principal components (PC) 1 through 10; and 5. eMERGE site. We performed logistic

regression-based association analyses for the case/control binary phenotype (general depression versus control) with the additive genotypic model of SNP genotypes coded as 0, 1, or 2 copies of the effective alleles using PLINK 2.0 [112]. The regional LD plot of the index SNP was created using the LDassoc web-based tool [113]. For ancestry-specific GWAS analysis (European, African American, and Asian ancestry), we used ancestry-specific PCs as covariates and kept other covariates unchanged. Following the initial stratified analyses, an additional logistic regression-based association analysis was performed in the European sample using the index SNP as a covariate to determine whether this SNP was truly driving the risk association.

7.3.5 *HLA* imputation

The *HLA* genotype imputation is done by implementing the HIBAG [114] R package with GRCH37 references with default parameters. We obtained three MHC-I class molecules (*HLA-A, HLA-B, HLA-C*) and four MHC-II class molecules (*HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPB1*).

7.3.6 *HLA* molecule association study

We narrowed down the *HLA* region and tried to identify if there are links between depression phenotypes and specific *HLA* molecules. The eMERGE I-III genotyped data contains imputed *HLA* genotypes using the HIBAG software for around 100k samples. We used all available samples from the eMERGE I-III genotyped patients to perform *HLA*-association analysis to identify *HLA* molecules (including class I and class II molecules) associated with depression susceptibility. We used logistic regressions to evaluate the association between each available *HLA* molecule and depression cases, controlling for age, sex, sites, and genetic ancestry.

7.3.7 Genetic correlation analysis

Genetic correlations are calculated using the tool LDSC [115]. Genetic correlation is a way of measuring how genetically correlated two different cohorts are, given the summary statistics of GWAS data. We computed the genetic correlation between all depression samples, major depression samples, non-major depression samples, and a large meta-analysis of major depression cohorts [97]. We use this method to validate our EHR-derived depression phenotypes and demonstrate the high genetic similarity between an existing large meta-analysis cohort and our cohort.

7.4 Results

7.3.1 GWAS analysis of combined ancestry

In total, 11,532 cases and 39,631 controls are included in the combined ancestry GWAS analysis. We observed minor differences in the age distribution between cases and controls in both the combined cohort (students' t-test = -2.67, P = 7.47e-3) and the European cohort (students' t-test = -3.43, P = 5.88e-4). In addition, we found a higher median body-mass index (BMI) (students' t-test, P = 9.42e-192) and a higher proportion of females in depression cases (odds = 2.04, P = 2.05e-234, combined ancestry; odds = 2.03, P = 1.92e-196, European cohort). This gender discrepancy is consistent with the well-known female predominance in the prevalence of with the current documentation and understanding of depression [116][117][118]. Researchers explored social and household factors in animal models to explain the discrepancies in depression rates in gender [117]. Existing opposite-sex twin research excluding genetic factors identified different sensitivity levels between women and men in social factors such as interpersonal relationships and goal-oriented factors [119].

In the combined ancestry analysis (Figure 7.1), we identified two known genes, *PHF5A* and *KCNG2* (with suggestive P < 1e-5, Table 7.1) that were previously are associated with depression (in GWAS of including unipolar depression, major depressive disorder, and bipolar disorder, derived queried from the GWAS

catalog) in both European and combined ancestry analysis [120]. Given that our sample size is limited compared to most large depression cohort GWAS studies, re-identifying existing genes supports the validity and practical use of the EHR-based depression phenotyping algorithm [102,121]. In African ancestry GWAS (n = 3950), we also re-identified three genes, including *SGCZ*, *ASIC2*, and *ZC3H7A* (Table 7.1) with suggestive P < 1e-5 that are associated with depression [102,122–124]. For Asian ancestry, limited by a small sample size (n = 583), we did not re-identify any known associations.

Figure 7.1 Manhattan plot of the combined ancestry GWAS analysis.

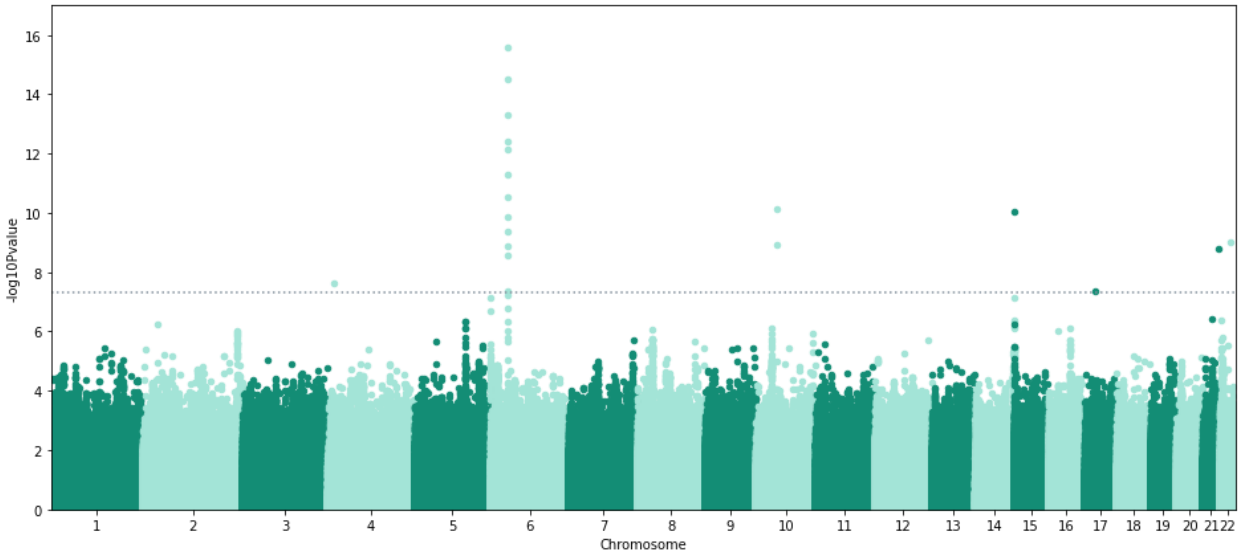


Table 7.1 Re-identified genes that are associated with depression from GWAS catalog

P<1e-5 in bold text

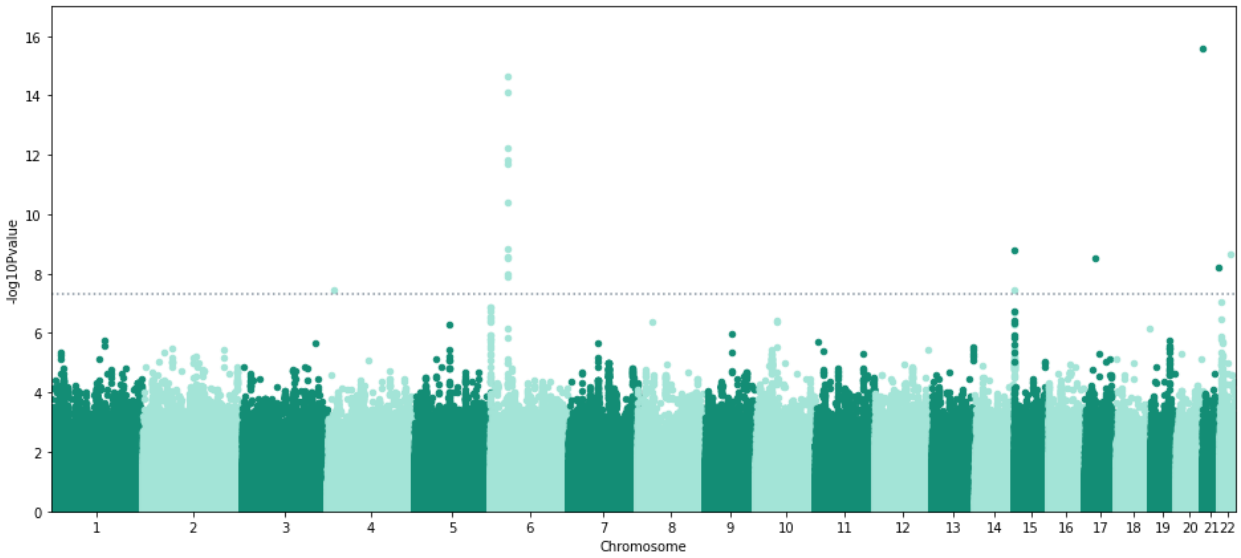
Gene	Reported traits	SNP (in this study)	P-value (Combined)	P-value (European)	P-value (African)	P-value (Asian)
PHF5A	Unipolar depression [125]	rs11705068	P=9.90e-10	P=2.28e-9	P=0.057	P=0.45
KCNG2	Unipolar depression [121,126],	rs35615281	P=1.30e-5	P=7.00e-7	P=0.84	P=0.26

	Major depressive disorder [127]					
ASIC2	Unipolar depression [102,122]	rs16968234	P=1.96e-4	P=0.27	P=8.80e-7	NA (constant allele)
SGCZ	Unipolar depression [102,123], Depressive symptom measure [123],	rs1383411	P=0.027	P=0.48	P=3.11e-7	P=0.43
ZC3H7A	Unipolar depression [124]	rs11644981	P=0.01	P=0.16	P=2.33e-6	P=0.88

7.3.2 GWAS analysis of European ancestry

The European ancestry GWAS analysis contains 9730 cases and 32,785 controls, which make up the majority of combined ancestry analysis. Besides identifying previously known associations, we also identified two new loci in the combined and European ancestry analysis that had not been reported previously in both the combined and European ancestry analysis (Figure 7.2).

Figure 7.2 Manhattan plot of the European ancestry GWAS analysis.



In the following analysis, we focus on this strong association between the depression phenotype and several within the human leukocyte antigen (*HLA*) region (Figure 7.2) in the European cohort. In the European ancestry analysis, 18 SNPs reached genome-wide significance (P < 5e-8), and 11 of them (top 3 listed as rs202207567, P=2.34e-15; rs28772724, P=8.1e-15; rs114031016, P=6.03e-15) fell in between the *HLA-DRB5* and *HLA-DRB1* genes, which belong to the MHC Class II region. Detailed summary statistics for the leading SNPs within the MHC Class II region are available in Table 7.2.

Table 7.2. Genome-Wide significant SNPs summary statistics on MHC-II region

Chr	SNP	Ref	Alt	BP(hg19)	MAF	Logistic European P-value OR (95% CI)	Logistic Combined P-value OR (95% CI)
6	rs202207567	C	T	32512533	0.13 (T)	**2.33e-15** **0.80 (0.68-0.93)**	**2.55e-16** **0.81 (0.70-0.94)**
6	rs28772724	G	T	32509357	0.14 (T)	**8.19e-15** **0.80 (0.69-0.94)**	**3.23e-15** **0.82 (0.72-0.94)**

6	rs114031016	C	T	32520035	0.13 (T)	**6.03e-13** **0.81 (0.70-0.95)**	**4.73e-14** **0.83 (0.72-0.95)**
6	rs113568276	G	A	32513127	0.13 (A)	**1.53e-12** **0.82 (0.70-0.95)**	**4.00e-13** **0.83 (0.72-0.95)**
6	rs111365964	T	G	32517646	0.13 (G)	**1.93e-12** **0.82 (0.70-0.95)**	**7.47e-13** **0.83 (0.73-0.94)**
6	rs76965357	T	C	32509778	0.15 (G)	**4.13e-11** **0.83 (0.73-0.96)**	**5.33e-12** **0.85 (0.75-0.96)**
6	rs112587701	T	A	32514041	0.16 (T)	**1.43e-9** **0.85 (0.74-0.97)**	**1.38e-10** **0.86 (0.76-0.97)**

The linkage disequilibrium (LD) plot of the index SNP (rs202207567, color in blue, Figure 7.3) and surrounding SNPs showed a red color gradient indicating the regulatory potential of each SNP, annotated by FORGEdb [128]. The third SNP rs114031016 and 5th SNP rs111365964 showed a high potential for regulatory functions, scoring 7 and 8, respectively. Additionally, searching in the Genotype-Tissue Expression (GTEx) portal, we found multiple tissue expression quantitative trait loci (eQTL) associated with four leading SNPs (rs113568276, rs112587701, rs28772724, rs111343881). Various MHC-II class molecule eQTLs (*HLA-DQA2, HLA-DRB6, HLA-DQA1, HLA-DRB1, HLA-DQB2, HLA-DQB1*) are highly significantly associated with these 4 SNPs (Figure 7.4). We colored the tissue type by Whole Blood versus the other, as most MHC-II class molecules are expressed by antigen-presenting cells (APCs) circulating in the blood. We found that the strongest P-value eQTLs are usually the Whole Blood tissue type in the GTEx portal, supporting the immune regulation aspect of depression phenotypes. The second peak of SNPs within chromosome 14 has a leading one with p = 3.66e-8 (rs4774137), falling into a region of the *IGHV* genes. These are the immunoglobulin heavy chain variable region genes, crucial for recognizing foreign antigens and initiating immune responses. Further, We identified multiple eQTLs of this SNP (rs4774137) in the GTEx data portal. Together, we identified the *HLA* region and immunoglobulin heavy chain regions that are associated with depression and reached a genome-wide significance.

Figure 7.3 LD plot for the index SNP rs202207567 in European ancestry

The blue dot indicates the index SNP rs202207567 and other SNPs are colored in a red gradient indicating their R^2 (darker color indicates higher R^2) relative to the index SNP. The numerical values are calculated scores indicating the regulatory function of individual SNP, ranging from 0 ~ 10, with a higher value indicating a higher probability of having regulatory functions.

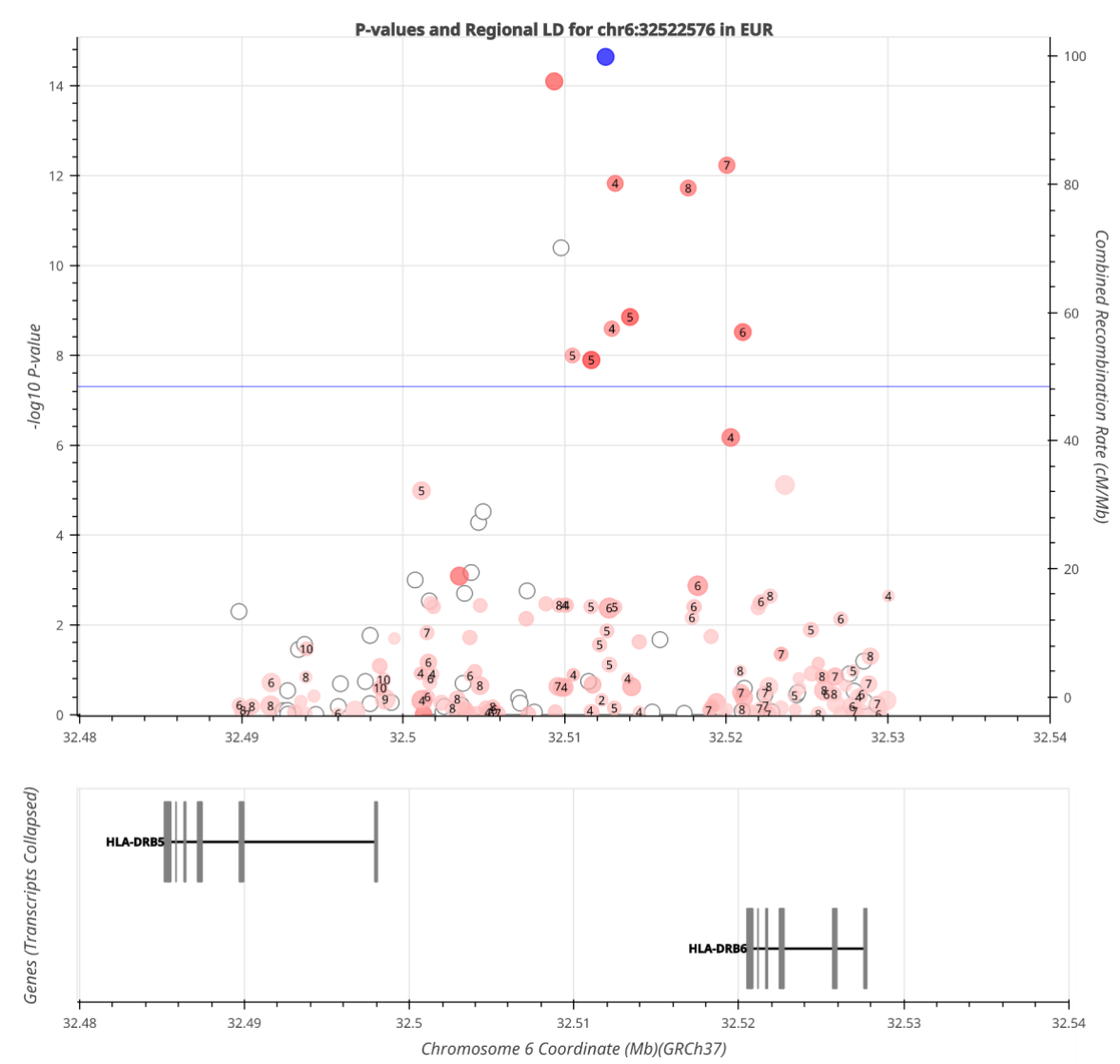
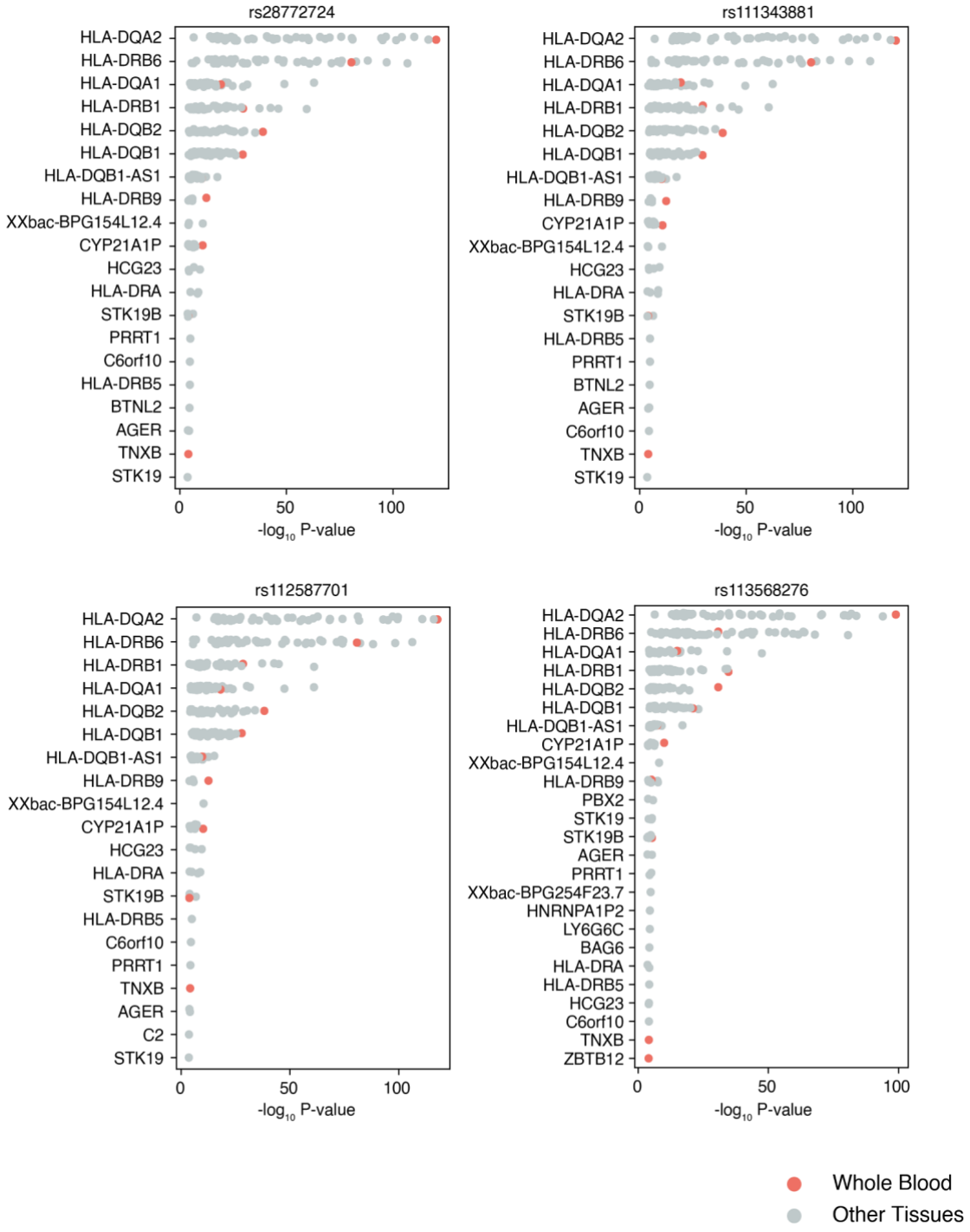


Figure 7.4 eQTLs for the four leading SNPs identified from GTEx

Each plot shows a tissue-wise eQTL plot with the x-axis indicating negative log10 p-values and the y-axis indicating corresponding genes. Each dot on the plot represents a specific tissue eQTL, with red dots emphasizing blood tissue, where most immune cells are circulating.



7.3.3 *HLA* association study on European ancestry

Based on the 11 SNPs we found in the MHC-II region, we performed association analysis on the European ancestry cohort, controlling for age, sex, median BMI, and sites, trying to identify *HLA* alleles associated with depression risk. We tested 236 *HLA*-alleles, including both MHC-I and MHC-II classes (*HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1,* and *HLA-DRB1*). 19 alleles showed a P-value < 0.05, but none were significant after Bonferroni multiple tests correction (Table 7.3). Interestingly, 13 out of the 19 alleles are MHC class I molecules, including 8 *HLA*-C alleles and 5 *HLA-B* alleles, with some appearing in high frequencies (*HLA*-B:0702 23.7%, *HLA*-C:0702 25.4%). The remaining 6 alleles are MHC class II molecules, including DQA1, DPB1, and DRB1. Though no significant alleles passed multiple test corrections, these results are suggestive, given that we have a relatively small sample size stratifying by each *HLA* molecule.

Table 7.3 Top hits of *HLA* association analysis summary statistics

Alleles	Coefficient	P-value	Fractions	Fractions_in_dep_cohort
DQA10601	0.500750	0.002938	0.010476	0.005341
C0801	1.179089	0.005699	0.004193	0.000716
C0403	1.933369	0.006234	0.000764	0.000269
B1516	0.882931	0.010447	0.007251	0.001134
DPB11901	0.292013	0.010545	0.009914	0.012473
DRB10803	0.464834	0.011019	0.005089	0.004595
C0304	0.085561	0.015342	0.136042	0.147913

DQB10504	0.584281	0.015477	0.002066	0.002477
B0702	-0.069634	0.016198	0.210997	0.237609
C0501	-0.080481	0.019164	0.143233	0.168561
C0702	-0.064044	0.022312	0.230240	0.254230
DRB11102	0.444519	0.025331	0.018419	0.003909
C0303	0.087626	0.033760	0.091582	0.106675
B4701	0.359414	0.038906	0.004205	0.005252
B4001	0.081300	0.040594	0.093111	0.116134
C1403	1.217580	0.042057	0.001111	0.000358
DQA10302	0.131090	0.044974	0.027605	0.034076
B3543	2.447207	0.048754	0.000621	0.000090
C1203	0.083981	0.049371	0.084344	0.097335

7.3.4 Genetic correlation validation of the phenotyping algorithm

To further assess the validation of our EHR-based depression phenotyping algorithm, we performed a genetic correlation analysis using LD score regression (LDSC) focusing on European ancestry, where we have sufficient samples [129]. The genetic correlation was made by comparing our algorithm to a large meta-analysis of 135,458 cases and 344,901 controls [97]. The summary statistics were downloaded from the PGC website. We found that the combined phenotype (all three types of depression, major depression

with psychosis, major depression, and non-major depression) had the highest genetic correlation with the meta-analysis (rg = 0.7419, sd =0.1347, p-values = 3.6158e-08). At the same time, major depression ranked second, with rg = 0.681, sd = 0.1474, and p-values = 3.827e-06. The least correlated phenotype is non-major depression, which had rg = 0.6551, sd = 0.1845, and p-values = 0.0004 (Table 7.4). These results indicate that our algorithm is also genetically valid and aligned very well with previously published large cohort analysis. Notably, the highest correlation between all depression (combined all three phenotypes) and the meta-analysis raised the potential that some existing major depression phenotyping algorithms might also include some non-major depression participants.

Table 7.4. LDSC regression for genetic correlation analysis

phenotype	rg	se	z-score	p
All depression	0.7419	0.1347	5.5087	3.6158e-08
Major depression	0.681	0.1474	4.6206	3.827e-06
Non-major depression	0.6551	0.1845	3.5505	0.0004

7.5 Conclusion

To this end, we have explored the secondary use of EHR from both unsupervised and supervised methods, demonstrating the potential of leveraging the EHR data to identify novel disease patterns and investigate disease etiology. In this chapter, we addressed the objective in Aim 3 (To perform a genome-wide association study (GWAS) for depression phenotype using rule-based phenotyping algorithm derived from

the EHR) and explored the genetic risk factor as a disease etiology for depression. We presented the use of a rule-based depression phenotyping algorithm from the EHR, constructed a cohort, and performed genome-wide scanning to identify genetic risk factors. This is the last chapter before the conclusion chapter (Chapter 8), which we marked as the last step of investigating disease patterns and disease etiology.

Depression is a mental health disorder characterized by persistent feelings of sadness, hopelessness, worthlessness, and a lack of interest in previously enjoyable activities. Depression can not only affect mood, behavior, and physical wellness but also interfere with daily activities, work, and relationships. Depression is typically diagnosed based on symptoms and medical history and is most known as a psychiatric disease. Only with the advances in genomic technology, the links between depression and genetic predisposition started to receive attention in the community [102],[97,130]. Yet, to carefully characterize a depression cohort is challenging. Much existing research relies on medical diagnosis codes or expertise-based screening processes. In this study, we implemented an EHR-based depression phenotyping algorithm, characterizing depression phenotypes across 12 sites in the eMERGE cohort, and conducted a GWAS. Our analysis validated the robustness of the EHR-based depression algorithm and yielded new results linking the depression phenotype with the immune system, expanding our current understanding of depression.

Our study has a few limitations. Firstly, the study's sample size is notably restricted (n = 42,515, EU), which falls short in contemporary genome-wide association studies (GWAS) for depression, which often encompass significantly larger cohorts to boost statistical power. Furthermore, we lose the granularity in characterizing the depression cohort, as the sample size is too small to segregate phenotypes into non-major depression, Major Depressive Disorder (MDD), and MDD with psychosis. Secondly, our phenotyping algorithm employed a stringent temporal criterion (180 days) to delineate events and encode outcomes as categorical values. This approach, while enhancing specificity, may overlook borderline cases and fail to capture the full spectrum of depression severity. A more optimal design is to incorporate the Patient Health

Questionnaire-9 (PHQ9) score and encoding outcome as continuous variables reflecting the likelihood or severity of depression, thereby augmenting the detection power.

Identification of MHC-II region SNPs and the *IGHV* region SNPs for the susceptibility of depression shed light linking these two phenotypes. The MHC-II region is one of the most polymorphic regions in the human genome. The MCH-II region is highly diverse because of the extensive genetic variation it exhibits, and the diversity is crucial for the recognition of various antigens, including pathogens like bacteria, viruses, and other foreign invaders. Genes within the MHC-II region encode cell surface proteins that present antigens to T cells, thus regulating the immune response, which is crucial for the immune system and plays a significant role in the adaptive immune response. The *IGHV* region genes are also polymorphic, which encode the variant domain of the heavy chain of immunoglobulin, also called antibodies. MHC-II proteins are mostly expressed on antigen-presenting cells (APCs) such as dendritic cells, macrophages, and B cells. These cells are involved in capturing, processing, and presenting antigens to helper T cells (CD4+ T cells). IGHVs are most functioning in activated B cells during the production of antibodies. Though the two regions are responsible for different functions, they all link to the immune response against pathogens.

There are previously identified genetic loci in the *HLA* region and depression phenotypes [131],[132],[122]. Besides, an early meta-analysis of clinical data revealed a decreased white blood cell count in depression patients [100]. Thus, existing evidence points to the strong association between immune dysfunction and depression phenotypes. However, it is still unclear the mechanisms and functional links between depression and the immune system. One explanation for this is that the wide range of symptoms of depression might directly be a reflection of immune dysregulation, given the genetic links, rendering the depression phenotype as a result of immunity. Another aspect would focus on the psychiatric part of depression, indicating that the occurrences of depression symptoms and progressions would lead to a series of immune function dysregulation and chronic inflammation. Or it could be simply bi-directional.

Together, our study identified genetic variants from the MHC-II region and the *IGVH* region that confer the susceptibility of depression in European ancestry, using an EHR-based phenotyping algorithm designed by the eMERGE consortium. Our analysis links depression with immune functions in genetics, supporting previously identified chronic inflammation features of depression and shedding new light on understanding the mechanisms of depression. Future research is needed for more deterministic causal inferences on these two components. It is becoming more clear that depression is not only a psychiatric disease, it has not only a complicated and wide range of symptoms but also affects multiple systems of the body, including the immune system.

In summary, this chapter concludes by completing the individual aims and demonstrating methods for studying novel disease patterns and disease etiology. In the next chapter (Chapter 8), we will summarize this work and briefly discuss future research directions.

Chapter 8: Conclusion

This chapter summarizes the primary contributions of this work, including the presented methodologies and empirical findings. It also presents future directions for continuing this research and plans for maximizing the impact of the resources created.

8.1 Summary

In this work, we explored the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. We split the objectives in each aim, 1) To develop patient representation learning in EHR data using an unsupervised machine learning approach; 2) To identify comorbidity patterns and progression trajectory variations using the longitudinal patient embedding vectors; 3) To perform a genome-wide association study (GWAS) for depression phenotype using rule-based phenotyping algorithm derived from the EHR. In this section, we will discuss each aim and their unique contributions, highlighting their significance not only to our objectives but also to the broader field.

The main contribution of this work will be illustrated in three unique aspects. The first one is the design of a novel neural network model architecture that achieved robust and outstanding performance in several downstream tasks. This neural network utilizes creative embeddings of disease onset frequency, calculated based on the eMERGE dataset, including 100k participants. The frequency-based embeddings are different from the commonly used semantic embeddings of diagnosis or procedure vocabularies, as we reason that the co-occurrences of disease and procedures matter more than the semantics. Due to this innovative approach, our model exhibited the power of multi-purposing, exhibiting excellent performances across different downstream tasks.

The second major contribution of this work is the identification of patient heterogeneities under certain well-defined phenotypes. Since the discovery of complex diseases, researchers have become aware that complicated interactions among genetics, behaviors, and environmental factors together determine the risk for certain diseases or phenotypes. However, nowadays it has become more clear that complex diseases are even more sophisticated, and patients vary in a wide degree of spectrum. For example, SLE patients can have their disease induced by certain drugs or exposure to chemical toxins, exhibiting at different levels of severity with distinct comorbidities (such as renal disease or cataracts). A comprehensive understanding of these discrepancies within a defined phenotype can further facilitate and achieve efficient precision medicine and personalized treatment. Our models unfolded the internal heterogeneities using SLE and CRC as two examples, demonstrating not only a static moment but a difference in the progression trajectory.

The third major contribution of this work is the demonstration of genetic studies using EHR-derived phenotyping algorithm, as a secondary use of EHR data retrospectively. To efficiently perform large genomic studies as a secondary use of the EHR data, accurate phenotyping algorithms are essential. We examined the depression phenotyping algorithm developed jointly by Kaiser Permanente and the University of Washington through genetics studies. Through GWAS analysis, we reproduced known disease-gene connections, meanwhile yielding novel associated SNPs within the *HLA* and *IGVH* region. We also observed a high genetic correlation between a large meta-analysis cohort and the cohort produced by our depression phenotyping algorithm, as extra validation of our phenotyping algorithm.

Our model has a few notable limitations. First, our model only included diagnosis and procedure codes as embedding building blocks, lacking medications, lab values, observations, and clinical notes due to the limitation of data sources. Without these variables, our model might lose certain meaningful information and limit the downstream analysis on medications, labs, etc. Besides, we used phecodes as surrogate phenotypes. Though phecodes have demonstrated their efficiency in large-scale EHR-based genetic studies, they might lack granularity and not be appropriate for some complicated phenotypes, such as

depression[133]. Last, our patient data drawn from the eMERGE consortium might contain potential ascertainment bias during patient recruitment, meaning that there might be population structures that can't represent the general population of the United States. However, on the other hand, with only diagnosis and procedure codes available, our model still demonstrated great performances in several downstream analyses, such as bulk phenotyping, disease forecasting, comorbidity pattern study, and progression analysis. This is not surprising, as diagnoses and procedures are the most crucial information within the EHR for many downstream tasks. Though phecodes are new and still in development, there is evidence that phecodes can reproduce genetic findings and serve as a great proxy for phenotypes. To adjust the potential biases caused by individual sites, we always included sites as covariates in our statistical analysis. Most importantly, we demonstrated the external validity of our model using the UW dataset and exhibited the robustness of performances with experimentally reproduced stable disease patterns invariant of cohorts.

In summary, this work demonstrated the potential of the secondary use of EHR, which has drawn huge attention in recent decades. We touched on several aspects of the secondary use of EHR, including phenotyping patients, predicting health outcomes, identifying disease subgroups, and trajectory (progression) analysis, through an unsupervised patient representation learning method. We then illustrated the use of an EHR-derived rule-based depression phenotyping algorithm to conduct a GWAS study, which further paved the road for the use of the EHR to unleash large populational-level genetic studies.

8.1 Future work

In this work, we explored the secondary use of EHR from both unsupervised and supervised methods, exploring the potential of utilizing the EHR data to identify novel disease patterns and investigate disease etiology. In this section, we will discuss future research regarding machine learning methods (regarding Aim 1, to develop patient representation learning in EHR data using an unsupervised machine learning approach and Aim 2, to identify comorbidity patterns and progression trajectory variations using the

longitudinal patient embedding vectors); and rule-based (Aim 3, to perform a genome-wide association study (GWAS) for depression phenotype using rule-based phenotyping algorithm derived from the EHR) methods, respectively.

In the latter part of Aim 1, we show that the advantages of the machine learning-based method lie in automation and massive parallel tasking ability. We demonstrated the use of patient embeddings in high-throughput phenotyping and disease onset prediction automatically without the interference of human labor. Moving forward, we show that our fine-tuned model using SBERT can perform clustering analysis to identify novel disease patterns that are potentially clinically important (e.g. hospital mortality). These tasks can be vastly benefited from automated processes using machine learning tools.

Regarding Aim 1, for future studies, researchers can focus on each of these single tasks to refine the model and adopt novel methods from computer science and deep learning fields to further improve the model performance and generalizability across different EHR systems. More practically, researchers can further investigate the health outcome of interest and use the embedding method to analyze discrepancies in health outcomes and identify clinically actionable items that benefit the patients.

For Aim 2, the analyses revealed complex internal heterogeneity of patients within a predefined phenotype/disease. In the future, researchers can carry out cross-sectional or even randomized control studies to rule out confounding variables and identify genuine connections between the heterogeneity of patients and important clinical outcome

In Aim 3, we utilized the rule-based phenotyping algorithm to create a refined depression cohort using the EHR data. For rule-based (expert-driven) methods and genetics, we propose two promising research directions. First, domain knowledge can support feature engineering for phenotyping, as traditional rule-based approaches often outperform machine learning in complex phenotyping tasks. By integrating domain

expertise, future studies can develop advanced feature engineering techniques to enhance phenotyping accuracy and provide complementary support for machine learning algorithms. Second, the extensive phenotypes derived from EHR data can facilitate the development and validation of polygenic risk scores (PRS). Given that many diseases are influenced by multiple genetic variants, these phenotypes could significantly advance the understanding of complex diseases and their genetic underpinnings.

References

1. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. J Am Med Inform Assoc. 2017;24: 1142–1148.

2. Jha AK. Meaningful use of electronic health records: the road ahead. JAMA. 2010;304: 1709–1710.

3. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine Learning Approaches for Electronic Health Records Phenotyping: A Methodical Review. medRxiv. 2022. p. 2022.04.23.22274218. doi:10.1101/2022.04.23.22274218

4. Bai T, Chanda AK, Egleston BL, Vucetic S. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. BMC Med Inform Decis Mak. 2018;18: 15–25.

5. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. [cited 20 Nov 2023]. Available: https://ieeexplore.ieee.org/abstract/document/8395074

6. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23: 1046–1052.

7. Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, et al. Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. Am J Hum Genet. 2016;99: 791.

8. Peissig PL, Santos Costa V, Caldwell MD, Rottscheit C, Berg RL, Mendonca EA, et al. Relational machine learning for electronic health record-driven phenotyping. J Biomed Inform. 2014;52: 260–270.

9. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine Learning Approaches for Electronic Health Records Phenotyping: A Methodical Review. medRxiv. 2022. p. 2022.04.23.22274218. doi:10.1101/2022.04.23.22274218

10. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018;1: 53–68.

11. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016;6: 26094.

12. Landi I, Glicksberg BS, Lee H-C, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. npj Digital Medicine. 2020;3: 1–11.

13. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. [cited 10 Nov 2023]. Available: https://ieeexplore.ieee.org/abstract/document/6472238?casa_token=5tJhmzs0ewcAAAAA:-

mQy65GfiYtgu09f-3MOYg9rDXyxyloosCnoE-NYwRkokfVTfM7ABHBqF1xXqv-eqFze6LdRyQ

14. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. J Biomed Inform. 2020;102: 103364.

15. Sathe NA, Xian S, Mabrey FL, Crosslin DR, Mooney SD, Morrell ED, et al. Evaluating construct validity of computable acute respiratory distress syndrome definitions in adults hospitalized with COVID-19: an electronic health records based approach. BMC Pulm Med. 2023;23: 292.

16. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2012;20: 117–121.

17. The Role of Domain Knowledge in Automating Medical Text Report Classification. J Am Med Inform Assoc. 2003;10: 330–338.

18. EHR-based phenotyping: Bulk learning and evaluation. J Biomed Inform. 2017;70: 35–51.

19. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13: 395–405.

20. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. Pediatrics. 2014;133: e54–e63.

21. Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015 [cited 22 Nov 2023]. doi:10.1126/scitranslmed.aaa9364

22. Zhang X, Chou J, Liang J, Xiao C, Zhao Y, Sarva H, et al. Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. Sci Rep. 2019;9: 1–12.

23. Becker F, Smilde AK, Acar E. Unsupervised EHR-based phenotyping via matrix and tensor decompositions. Wiley Interdiscip Rev Data Min Knowl Discov. 2023;13: e1494.

24. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016;6: 26094.

25. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. KDD. 2017;2017: 787–795.

26. Impact of integrating genomic data into the electronic health record on genetics care delivery. Genetics in Medicine. 2022;24: 2338–2350.

27. Wolford BN, Willer CJ, Surakka I. Electronic health records: the next wave of complex disease genetics. Hum Mol Genet. 2018;27: R14–R21.

28. Shoenbill K, Fost N, Tachinardi U, Mendonca EA. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc. 2013;21: 171–180.

29. Returning integrated genomic risk and clinical recommendations: The eMERGE study. Genetics in Medicine. 2023;25: 100006.

30. Bush WS, Crosslin DR, Owusu-Obeng A, Wallace J, Almoguera B, Basford MA, et al. Genetic

variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. Clinical Pharmacology & Therapeutics. 2016;100: 160–169.

31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 2017. Available: http://arxiv.org/abs/1706.03762

32. Gijsen R, Hoeymans N, Schellevis FG, Ruwaard D, Satariano WA, van den Bos GA. Causes and consequences of comorbidity: a review. J Clin Epidemiol. 2001;54: 661–674.

33. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4: 13.

34. Doshi H, Solli E, Elze T, Pasquale LR, Wall M, Kupersmith MJ. Unsupervised Machine Learning Identifies Quantifiable Patterns of Visual Field Loss in Idiopathic Intracranial Hypertension. Transl Vis Sci Technol. 2021;10: 37.

35. Ash JS, Gorman PN, Lavelle M, Payne TH, Massaro TA, Frantz GL, et al. A cross-site qualitative study of physician order entry. J Am Med Inform Assoc. 2003;10: 188–200.

36. Prevalence and clinical course of depression: A review. Clin Psychol Rev. 2011;31: 1117–1125.

37. Stein M, Miller AH, Trestman RL. Depression, the Immune System, and Health and Illness: Findings in Search of Meaning. Arch Gen Psychiatry. 1991;48: 171–177.

38. Kook AI, Mizruchin A, Odnopozov N, Gershon H, Segev Y. Depression and immunity: the biochemical interrelationship between the central nervous system and the immune system. Biol Psychiatry. 1995;37: 817–819.

39. Leonard BE, Song C. Stress and the immune system in the etiology of anxiety and depression. 1996.

40. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc. 2013;20: e206–11.

41. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. pp. 1532–1543.

42. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. Available: http://arxiv.org/abs/1301.3781

43. Brunette ES, Flemmer RC, Flemmer CL. A review of artificial intelligence. [cited 14 Jul 2024]. Available: https://ieeexplore.ieee.org/abstract/document/4804025

44. Rojas R. The Backpropagation Algorithm. Neural Netw. 1996; 149–182.

45. Backpropagation neural networks: A tutorial. Chemometrics Intellig Lab Syst. 1993;18: 115–155.

46. Zhai J, Zhang S, Chen J, He Q. Autoencoder and Its Various Variants. [cited 14 Jul 2024]. Available: https://ieeexplore.ieee.org/abstract/document/8616075

47. Autoencoder for words. Neurocomputing. 2014;139: 84–96.

48. Wang X, McCallum A, Wei X. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. [cited 14 Jul 2024]. Available: https://ieeexplore.ieee.org/abstract/document/4470313

49. Sidorov G. Syntactic n-grams in Computational Linguistics. Springer; 2019.

50. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl. 2018;78: 15169–15211.

51. Lenci A, Sahlgren M. Distributional Semantics. Cambridge University Press; 2023.

52. Boleda G. Distributional Semantics and Linguistic Theory. Annu Rev Appl Linguist. 2020;6: 213–234.

53. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine. 2021;4: 1–13.

54. Staudemeyer RC, Morris ER. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. 2019. Available: http://arxiv.org/abs/1909.09586

55. Schmidt RM. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. 2019. Available: http://arxiv.org/abs/1912.05911

56. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Medical Informatics. 2020;8: e17984.

57. Nuthakki S, Neela S, Gichoya JW, Purkayastha S. Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks. 2019. Available: http://arxiv.org/abs/1912.12397

58. Lybarger K, Ostendorf M, Thompson M, Yetisgen M. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. J Biomed Inform. 2021;117: 103761.

59. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019;36: 1234–1240.

60. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. Available: http://arxiv.org/abs/1810.04805

61. Casey G. The BRCA1 and BRCA2 breast cancer genes. Curr Opin Oncol. 1997;9: 88.

62. Verma A, Damrauer SM, Naseer N, Weaver J, Kripke CM, Guare L, et al. The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. Journal of Personalized Medicine. 2022;12: 1974.

63. Zheng NS, Stone CA, Jiang L, Shaffer CM, Eric Kerchberger V, Chung CP, et al. High-throughput framework for genetic analyses of adverse drug reactions using electronic health records. PLoS Genet. 2021;17: e1009593.

64. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing Rule-Based and Deep Learning Models for Patient Phenotyping. 2017. Available: http://arxiv.org/abs/1703.08705

65. Choi E, Xiao C, Stewart W, Sun J. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. Adv Neural Inf Process Syst. 2018;31. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/934b535800b1cba8f96a5d72f72f1611-Paper.pdf

66. Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A. Deep Learning Techniques for Biomedical and Health Informatics. Springer Nature; 2019.

67. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018;22: 1589–1604.

68. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. Adv Neural Inf Process Syst. 2017;30. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

69. Tang AS, Oskotsky T, Havaldar S, Mantyh WG, Bicak M, Solsberg CW, et al. Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. Nat Commun. 2022;13: 1–15.

70. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9: 2579–2605.

71. Zhu Z, Yin C, Qian B, Cheng Y, Wei J, Wang F. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. [cited 25 Nov 2023]. Available: https://ieeexplore.ieee.org/document/7837899

72. Pranjul Yadav University of Minnesota-Twin Cities, MN, USA, Michael Steinbach University of Minnesota-Twin Cities, MN, USA, Vipin Kumar University of Minnesota-Twin Cities, MN, USA, Gyorgy Simon University of Minnesota-Twin Cities, MN, USA. Mining Electronic Health Records (EHRs): A Survey. In: ACM Computing Surveys [Internet]. [cited 25 Nov 2023]. Available: https://dl.acm.org/doi/10.1145/3127881

73. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 2014;30: 2375–2376.

74. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nature Protocols. 2019;14: 3426–3444.

75. Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc. 2013;20: e253–9.

76. Gehan MA, Kellogg EA. High-throughput phenotyping. Am J Bot. 2017;104: 505–508.

77. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. Machine Learning for Healthcare Conference. PMLR; 2016. pp. 301–318.

78. Liu J, Zhang Z, Razavian N. Deep EHR: Chronic Disease Prediction Using Medical Notes. Machine

Learning for Healthcare Conference. PMLR; 2018. pp. 440–464.

79. Dammacco R. Systemic lupus erythematosus and ocular involvement: an overview. Clin Exp Med. 2018;18: 135–149.

80. Alderaan K, Sekicki V, Magder LS, Petri M. Risk factors for cataracts in systemic lupus erythematosus (SLE). Rheumatol Int. 2015;35: 701–708.

81. A national study of the complications of lupus in pregnancy. Am J Obstet Gynecol. 2008;199: 127.e1–127.e6.

82. Baer AN, Witter FR, Petri M. Lupus and pregnancy. Obstet Gynecol Surv. 2011;66: 639–653.

83. O'Neill TJ, Nguemo JD, Tynan A-M, Burchell AN, Antoniou T. Risk of Colorectal Cancer and Associated Mortality in HIV: A Systematic Review and Meta-Analysis. J Acquir Immune Defic Syndr. 2017;75: 439.

84. Newnham A, Harris J, Evans HS, Evans BG, Møller H. The risk of cancer in HIV-infected people in southeast England: a cohort study. Br J Cancer. 2005;92: 194.

85. Cooksley CD, Hwang LY, Waller DK, Ford CE. HIV-related malignancies: community-based study using linkage of cancer registry and HIV registry data. Int J STD AIDS. 1999;10. doi:10.1258/0956462991913574

86. Chen C-H, Chung C-Y, Wang L-H, Lin C, Lin H-L, Lin H-C. Risk of cancer among HIV-infected patients from a population-based nested case–control study: implications for cancer prevention. BMC Cancer. 2015;15. doi:10.1186/s12885-015-1099-y

87. Fehringer G, Kraft P, Pharoah PD, Eeles RA, Chatterjee N, Schumacher FR, et al. Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. Cancer Res. 2016;76: 5103–5114.

88. Rashkin SR, Graff RE, Kachuri L, Thai KK, Alexeeff SE, Blatchins MA, et al. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat Commun. 2020;11: 1–14.

89. Hawkins ML, Blackburn BE, Rowe K, Snyder J, Deshmukh VG, Newman M, et al. Endocrine and Metabolic Diseases Among Colorectal Cancer Survivors in a Population-Based Cohort. J Natl Cancer Inst. 2019;112: 78–86.

90. Barone M, Lofano K, De Tullio N, Licino R, Albano F, Di Leo A. Dietary, Endocrine, and Metabolic Factors in the Development of Colorectal Cancer. J Gastrointest Cancer. 2011;43: 13–19.

91. Kenzik KM, Balentine C, Richman J, Kilgore M, Bhatia S, Williams GR. New-Onset Cardiovascular Morbidity in Older Adults With Stage I to III Colorectal Cancer. J Clin Oncol. 2018 [cited 20 Nov 2023]. doi:10.1200/JCO.2017.74.9739

92. Risk of Arterial Thromboembolism in Patients With Cancer. J Am Coll Cardiol. 2017;70: 926–938.

93. Coghill AE, Suneja G, Rositch AF, Shiels MS, Engels EA. HIV Infection, Cancer Treatment Regimens, and Cancer Outcomes Among Elderly Adults in the United States. JAMA Oncol. 2019;5: e191742–e191742.

94. Wang J, Wu X, Lai W, Long E, Zhang X, Li W, et al. Prevalence of depression and depressive symptoms among outpatients: a systematic review and meta-analysis. BMJ Open. 2017;7: e017173.

95. Penninx BW, Milaneschi Y, Lamers F, Vogelzangs N. Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile. BMC Med. 2013;11: 1–14.

96. Sullivan PF, Neale MC, Kendler KS. Genetic Epidemiology of Major Depression: Review and Meta-Analysis. Am J Psychiatry. 2000 [cited 1 Nov 2023]. doi:10.1176/appi.ajp.157.10.1552

97. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50: 668–681.

98. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012;90: 7.

99. Immune dysregulation in depression: Evidence from genome-wide association. Brain, Behavior, & Immunity - Health. 2020;7: 100108.

100. Herbert TB, Cohen S. Depression and immunity: a meta-analytic review. Psychol Bull. 1993;113: 472–486.

101. Stress, depression, the immune system, and cancer. Lancet Oncol. 2004;5: 617–625.

102. Levey DF, Stein MB, Wendt FR, Pathak GA, Zhou H, Aslan M, et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. Nat Neurosci. 2021;24: 954–963.

103. Levey DF, Stein MB, Wendt FR, Pathak GA, Zhou H, Aslan M, et al. GWAS of Depression Phenotypes in the Million Veteran Program and Meta-analysis in More than 1.2 Million Participants Yields 178 Independent Risk Loci. medRxiv. 2020. p. 2020.05.18.20100685. doi:10.1101/2020.05.18.20100685

104. Stanaway IB, Hall TO, Rosenthal EA, Palmer M, Naranbhai V, Knevel R, et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. Genet Epidemiol. 2019;43. doi:10.1002/gepi.22167

105. Sealock JM, Lee YH, Moscati A, Venkatesh S, Voloudakis G, Straub P, et al. Use of the PsycheMERGE Network to Investigate the Association Between Depression Polygenic Scores and White Blood Cell Count. JAMA Psychiatry. 2021;78: 1365–1374.

106. Depression. [cited 4 Dec 2023]. Available: https://phekb.org/phenotype/depression

107. Sheehan DV. Depression: underdiagnosed, undertreated, underappreciated. Manag Care. 2004;13: 6–8.

108. Mortality and life expectancy in persons with severe unipolar depression. J Affect Disord. 2016;193: 203–207.

109. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48: 1284–1287.

110. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48: 1443–1448.

111. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4: 7.

112. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575.

113. Machiela MJ, Chanock SJ. LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation. Bioinformatics. 2018. Available: https://academic.oup.com/bioinformatics/article-abstract/34/5/887/4124856

114. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG—HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2013;14: 192–200.

115. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. Nature Genetics. 2015;47: 1236–1241.

116. Depression in women: Understanding the gender gap. In: Mayo Clinic [Internet]. 29 Jan 2019 [cited 16 Oct 2023]. Available: https://www.mayoclinic.org/diseases-conditions/depression/in-depth/depression/art-20047725

117. Animal models of anxiety and depression: how are females different? Neurosci Biobehav Rev. 2001;25: 219–233.

118. Albert PR. Why is depression more prevalent in women? J Psychiatry Neurosci. 2015;40: 219.

119. Kendler KS, Gardner CO. Sex Differences in the Pathways to Major Depression: A Study of Opposite-Sex Twin Pairs. Am J Psychiatry. 2014;171: 426.

120. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2016;45: D896–D901.

121. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22: 343–352.

122. Howard DM, Adams MJ, Shirali M, Clarke T-K, Marioni RE, Davies G, et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. Nat Commun. 2018;9: 1470.

123. Fabbri C, Kasper S, Kautzky A, Bartova L, Dold M, Zohar J, et al. Genome-wide association study of treatment-resistance in depression and meta-analysis of three independent samples. Br J Psychiatry. 2019;214: 36–41.

124. Heinzman JT, Hoth KF, Cho MH, Sakornsakolpat P, Regan EA, Make BJ, et al. GWAS and systems biology analysis of depressive symptoms among smokers from the COPDGene cohort. J Affect Disord. 2019;243: 16–22.

125. Yu AQ, Wang J, Jiang ST, Yuan LQ, Ma HY, Hu YM, et al. SIRT7-Induced PHF5A

Decrotonylation Regulates Aging Progress Through Alternative Splicing-Mediated Downregulation of CDK2. Frontiers in cell and developmental biology. 2021;9. doi:10.3389/fcell.2021.710479

126. Mitchell BL, Campos AI, Whiteman DC, Olsen CM, Gordon SD, Walker AJ, et al. The Australian Genetics of Depression Study: New Risk Loci and Dissecting Heterogeneity Between Subtypes. Biol Psychiatry. 2022;92: 227–235.

127. Yao X, Glessner JT, Li J, Qi X, Hou X, Zhu C, et al. Integrative analysis of genome-wide association studies identifies novel loci associated with neuropsychiatric disorders. Transl Psychiatry. 2021;11: 69.

128. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22: 1790–1797.

129. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47: 291–295.

130. Mullins N, Lewis CM. Genetics of Depression: Progress at Last. Curr Psychiatry Rep. 2017;19: 1–7.

131. Thorp JG, Campos AI, Grotzinger AD, Gerring ZF, An J, Ong J-S, et al. Symptom-level modelling unravels the shared genetic architecture of anxiety and depression. Nat Hum Behav. 2021;5: 1432–1442.

132. Coleman JRI, Peyrot WJ, Purves KL, Davis KAS, Rayner C, Choi SW, et al. Genome-wide gene-environment analyses of major depressive disorder and reported lifetime traumatic experiences in UK Biobank. Mol Psychiatry. 2020;25: 1430–1446.

133. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One. 2017;12: e0175508.