PHENOTREE: Interactive Visual Analytics for Hierarchical Phenotyping from Large-Scale Electronic Health Records

Inci M. Baytas, Kaixiang Lin, Fei Wang, Anil K. Jain, *Life Fellow, IEEE*, and Jiayu Zhou

Electronic health records (EHRs) capture comprehensive patient information in digital form from a variety of sources. Increasing availability of EHRs has facilitated development of data and visual analytic tools for healthcare analytics, such as clinical decision support and patient care management systems. Many healthcare analytic tools are used to investigate fundamental problems, such as study of patient population, exploring complicated interactions among patients and their medical histories, and extracting structured phenotypes characterizing the patient population. In this paper, we propose PHENOTREE a novel data-driven, hierarchical, and interactive phenotyping tool, that enables physicians and medical researchers to participate in the phenotyping process of large-scale EHR cohorts. The proposed visual analytic tool allows users to interactively explore EHR cohorts, and generate, interpret, evaluate and refine phenotypes by building and navigating a phenotype hierarchy. Specifically, given a cohort or sub-cohort, PHENOTREE employs sparse principal component analysis (SPCA) to identify key clinical features that characterize the population. The clinical features provide a natural way to generate deeper phenotypes at finer granularities by expanding the phenotype hierarchy. To facilitate the intensive computation required for interactive analytics, we design an efficient SPCA solver based on variance reduced stochastic gradient technique. The benefits of our method are demonstrated by analyzing two different EHR patient cohorts, a public and a private dataset containing EHRs of $101,767$ and $223,076$ patients, respectively. Our evaluations show that PHENOTREE can detect clinically meaningful hierarchical phenotypes.

***Index Terms*—Electronic health records, interactive visual analytics, hierarchical phenotyping, sparse principal component analysis, data-driven phenotyping.**

I. INTRODUCTION

ELECTRONIC health records (EHRs) provide digital means to capture comprehensive patient information from a variety of data sources, such as inpatient/outpatient encounters, diagnostic records, medication history, medical images, lab test panels, etc. The adoption rate of EHR systems in the United States has significantly increased over the past decade, from 18% in 2001 to over 78% in 2013 [1]. As EHR systems become more prevalent and assimilate more information, it becomes challenging for physicians to draw clinical conclusions by analyzing raw EHR data. On the other hand, availability of vast amount of EHR data has given researchers an unprecedented opportunity to develop advanced healthcare analytic techniques for improving patient care. Development of visual and data analytic tools has greatly impacted many aspects of healthcare such as clinical care management and decision support applications.

One of the fundamental challenges in healthcare systems is to identify clinically important *phenotypes*[1] that characterize patient cohorts. With the increasing capability of computational modeling and availability of large medical data archives, it is now possible to develop data-driven approaches for *computational phenotyping* that utilize machine learning algorithms to infer phenotypes from complex historical medical records [2]–[5]. To increase the impact of computational phenotyping in clinical practice, the National Institute of Health Big Data to Knowledge (BD2K) Initiative [6] has sponsored the Center for Predictive Computational Phenotyping for developing advanced computational phenotyping techniques.

An impending challenge in computational phenotyping is

[1]The term phenotype is generally used to denote a composite of observable properties of an organism, as a result of sophisticated interactions between its genotype and environmental surroundings [2].

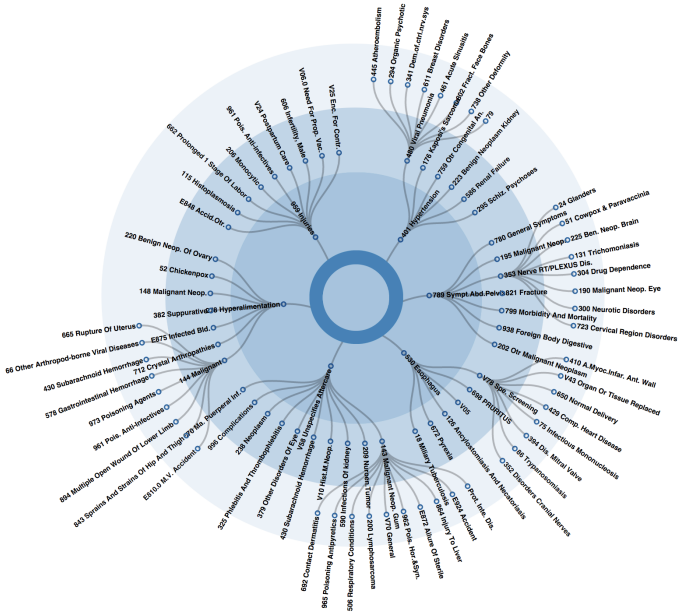


Fig. 1: An example of PHENOTREE that discovers hierarchical phenotypes in an EHR cohort. Starting from a large-scale EHR cohort, the proposed method applies sparse principal component analysis (SPCA) to identify the key medical features that characterize the cohort. These key features provides a natural way to identify subcohorts whose patients are associated with one of the features. The PHENOTREE then applies SPCA to the subcohorts to identify finer levels of granularity of phenotypes. Each node in this tree gives a structured phenotype and a stable subcohort characterized by this phenotype.

I. M. Baytas, K. Lin, A. K. Jain, and J. Zhou are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824 USA. Correspondence: jiayuz@msu.edu

F. Wang is with the Division of Health Informatics, Department of Healthcare Policy and Research, Cornell University, New York, NY.

to determine the granularity of phenotypes [7], [8]. While most existing computational phenotyping techniques capture coarse characteristics of the population and provide high-level descriptions of the stable subcohorts (i.e., flat phenotypes), it is often more interesting to explore deeper phenotypes of finer levels of granularity, as well as hierarchies within the phenotypes. Such hierarchical phenotypes are comparatively more complex than traditional flat phenotypes because the hierarchical phenotypes have complicated interactions among layered structures. Furthermore, because of the combinatorial nature of hierarchical phenotypes, a large number of hierarchical phenotypes are possible in the absence of adequate input from medical experts. The complexity and the scale of the hierarchical phenotypes require involvement of human experts, such as physicians and medical researchers, in phenotyping. To assist the human experts in interactively exploring, interpreting, evaluating and refining the phenotype hierarchy, development of visual analytic tools is essential. Although existing tools can be used to visualize phenotype hierarchies, they are not specifically designed for interactive phenotyping where the human experts need to build and navigate through the hierarchy during cohort exploration, frequently performing phenotyping on different subcohorts. Long processing times of existing phenotyping algorithms prohibit interactive analytics while analyzing large-scale EHR datasets.

To overcome the aforementioned challenges, we introduce PHENOTREE a novel interactive visual analytics tool for medical experts to perform efficient and effective hierarchical phenotyping for large patient cohorts. Specifically, PHENOTREE enables 1) interactive exploration of large-scale EHR cohorts, 2) hierarchical phenotype discovery by iterative application of sparse principal component analysis (SPCA), and 3) visualization of phenotypes at different levels of granularity as a tree structure that helps medical experts interpret, evaluate and adjust hierarchical phenotypes. Given cohort/subcohorts, PHENOTREE employs SPCA to identify key clinical features and then generates phenotypes based on these features. Patients associated with each key clinical feature are then grouped into individual subcohorts. Through this interactive process, PHENOTREE helps medical experts to identify sets of phenotypes and corresponding patient subcohorts at different granularities. Because existing solvers for SPCA have high computational complexities and therefore are not suitable for interactive user experience, we propose to adopt a convex formulation for sparse PCA and then apply a variance reduced stochastic gradient technique that achieves fast convergence rates to overcome the computational challenge of SPCA. We demonstrate the advantages of PHENOTREE by conducting experiments on two real-world EHR patient cohorts. We show that the computational efficiency of the proposed method significantly outperforms traditional SPCA solvers. We also discuss some interesting findings about hierarchical phenotypes obtained using the proposed interactive visual analytic technique.

The rest of this paper is organized as follows: In Section II we overview related work on data-driven phenotyping, sparse principal component analysis and stochastic proximal optimization, and visual analytics for EHR. We then present our proposed approach in Section III. We present and discuss our experimental results in Section IV and Section V concludes the paper.

II. RELATED WORK

This work builds upon three major research directions: data-driven phenotyping, sparse principal component analysis with stochastic proximal optimization and visual analytic tools for analyzing EHRs.

A. Data-Driven Phenotyping

The clinical data used in this study is electronic health records which comprise of ICD9 [2] diagnoses of each patient. Working with EHR can be quite challenging, since the data is generally very sparse and noisy. Therefore, more stable and robust ways to describe patients is a necessity. One way of representing patients is phenotyping. Phenotyping patients by using clinical records is a commonly studied problem in recent years. As it was discussed in [2], extracting phenotypic patterns of patients is an important task which can contribute to the development of personalized medicine.

One of the studies, where a data driven phenotyping framework was proposed by using a longitudinal electronic health data, is by Zhou *et al.* in [2]. They developed a phenotyping framework which called PACIFIER, and showed that the proposed approach improves predictive performance in two real world EHR cohorts. The main assumption in their study was that the medical features of EHR data can be mapped to a much lower dimensional latent space. Thus, each medical concept was assumed to be the combination of several observed medical features which were called as macro-phenotypes. The two formulations, called Individual Basis Approach and Shared Basis Approach, were developed to obtain a compact representation of the patients. As a summary, PACIFIER was proposed to perform temporal matrix completion via low- rank factorization. This paper suggests to densify the sparse EHR data by making use of the longitudinal information.

Deep learning has received significant attention and has now been applied to many areas. So it is not surprising that deep networks have also used for phenotyping. For instance, Che *et al.* proposed to use deep learning for the discovery and detection of characteristic patterns in clinical data in [9]. Some modifications to standard neural net training were done to be able to utilize deep learning for medical data. Their experiments on two real world health care data sets showed that neural networks can learn relevant features for medical applications. Che *et al.* stated that the proposed deep learning framework improves the multi-label classification performance such as predicting ICD9 codes. In this study, neural networks were trained on windows of multivariate clinical time series data. Time information was incorporated as in [2]. Existing domain knowledge was used as a prior which was transformed into a regularizer in the learning phase to deal with the

[2]The International Conference for the Ninth Revision of the International Classification of Diseases

limited data. Deep learning is currently used as an effective method for supervised problems, whereas our problem has an unsupervised flavor.

Another interesting study of phenotyping is [4] where a sparse non-negative tensor factorization method was proposed to obtain some phenotype candidates. The approach suggested in [4], called Marble, is based on decomposing the observed tensor into two terms as bias and interaction tensors. Phenotypes were defined by the interaction tensor without any human supervision. Ho *et al.* also defined the properties of an ideal phenotype as being representative of the complex interactions between several sources, being compact and comprehensible to medical professionals and ability of mapping to domain knowledge. According to this definition, each phenotype was suggested as a latent space and phenotyping was thought as a dimensionality reduction approach. This was how the relationship between phenotyping and the tensor factorization was established. Tensor factorization can also be thought as a dimensionality reduction method. Since the tensor-derived phenotypes introduced redundancy, a sparse non-negative tensor factorization was utilized to obtain concise phenotypes. Marble was proposed as a basis for an automated high-throughput phenotyping tool.

In this paper, we are dealing with an unsupervised problem with no prior information. In the same spirit, the problem of discovering the patterns in EHR data was modeled as an unsupervised problem in [10]. Marlin *et al.* addressed the temporal sparsity of EHR data by developing a probabilistic clustering model with an empirical prior distribution which was used to deal with the sparsity of data. The proposed approach was developed for multidimensional, sparse and uncertain physiological time series data collected from real world EHRs. Marlin *et al.* stated that their model can capture physiological patterns, and the clusters generated by the proposed model indicate clear differences in the path of different physiological variables. Some limitations of the proposed framework which are mostly related to time aspect of the health care data were also discussed.

B. Sparse Principal Component Analysis and Stochastic Proximal Optimization

Zou *et al.* proposed the sparse PCA for the first time in [11]. Authors posed the problem of learning sparse loading vectors as a regression problem with the lasso (elastic net) constraint. On the other hand, d'Aspremont *et al.* had a sparse PCA solution by utilizing semi-definite programming in [12]. Results showed that proposed approach helps to improve the sparsity of the solution. However, these approaches are not scalable in terms of number of samples. Journee *et al.* in 2010 introduced two types of sparse PCA algorithms namely single unit and block sparse PCA in [13]. Their formulations were based on maximizing a convex function on a compact set using ℓ_1 or ℓ_0 norms. Authors showed that their algorithms are faster when the objective function or feasible set are strongly convex. In both cases, large scale and high dimensional data is difficult to handle. In [14], a generalization of the inverse power method was derived by using constrained optimization problems with non-quadratic objective and constraints. An inverse power method was used for the sparse PCA and the spectral clustering.

Naikal *et al.*, 2011 [15] proposed a sparse PCA approach for informative feature selection. The experiments indicated that using sparse PCA improves the object recognition accuracy. Sparse PCA was formulated a semi-definite programming problem with augmented Lagrangian method. A more recent study [16] investigated sparse PCA with oracle property. Their proposed method has a family of estimators based on semi-definite relaxation of sparse PCA and the algorithm estimates k-dimensional principal subspace of a population matrix based on sample covariance matrix. Finally, a stochastic PCA algorithm with exponential convergence rate was proposed in [17]. Authors defined an efficient algorithm for inexpensive stochastic iterations and variance reduction which was suggested in [18]. Optimization scheme used in [17] requires strong convexity. Since PCA is a non-convex problem, authors utilized a different convergence analysis than the convergence analysis of [18].

In this paper, we utilize a stochastic approach which is advantageous for dealing with large number of samples compared to the studies summarized above. We also use the convex optimization formulation of finding the leading eigenvector, which was proposed in [19], to be able to exploit well defined convergence analysis of proximal stochastic method with variance reduction. In the literature, there are several proximal gradient based methods [20], [21] developed recently. One of the representative works is FISTA by Beck and Teboulle [20]. FISTA provided the fastest convergence rate among first order methods for full gradient descent. However, when the number of samples is very large, approaches using full gradient will not be scalable enough. Therefore, stochastic gradient methods are generally used in problems with large sample sizes. However, the problem of stochastic algorithms is the low convergence because of the high variance of the gradient.

Nitanda [22] proposed a variance reduction framework combined with Nesterov's acceleration method to mitigate the aforementioned drawback of traditional stochastic gradient methods. Another approach proposed in [23], studied a stochastic gradient method that provides exponential convergence rate. Furthermore, Johnson and Zhang also proposed a progressive variance reduced proximal gradient method in [18]. Variance reduction mechanism computes the gradient by making use of the full gradient and the estimated optimal point at each iteration. Strong convexity of the objective function was assumed to achieve a geometric convergence rate under expectation. Xiao and Zhang presented another variance reduced stochastic approach for proximal algorithms in [24]. A multi-stage scheme was proposed to reduce the variance of the stochastic gradient progressively. Similar to [18], strong convexity and the Lipschitz continuity were the basic assumptions.

C. Visual Analytics for EHR

EHRs contain huge amount of information and it is not straightforward to explore patterns and structures from raw

electronic records for medical experts by using conventional methods. Visualization helps the experts to understand, interpret, and discover undercover information in EHRs. Therefore, numerous studies have been done on developing methods for visual analysis of patients records. In one of the studies, Perer *et al.* proposed a system for mining and visualizing the frequent event sequences from EHR data in [25]. Developed system was called as Care Pathway Explorer, which presents a technique to mine frequent sequences from EHR patient traces and to visualize the mined patterns with an interactive user interface.

Gotz *et al.* also proposed a visualization technique with mining exploratory data within EHRs in [26]. Their method provided on-demand analysis of clinical sequence data and an interactive visual interface that users can retrieve patient cohorts easily. Wang *et al.* studied a visual analysis method to improve cohort studies of EHR for chronic kidney disease in [27]. Their proposed system, which allows users to make visual aggregation, provided an interactive visual mining interface to support explorative analysis of high dimensional EHR data.

Another study which proposed a standardized data analysis process for cohort studies by using EHR data is [28]. Huang *et al.* utilized an interactive approach to classify patients into different subsets, which are used for visualization along with user feedback. Authors developed a visually rich web-based application that can help physicians and researchers to comprehend and study patient cohorts over time.

III. Methodology

A. Electronic Health Records

Electronic health records consist of diagnostics information of patients collected over a period of time [29]. Diagnostic information in EHRS is generally coded according to ICD9, for each patient. Each ICD9 code corresponds to a specific diagnosis and every diagnosis belongs to a broader diagnosis group. For instance, ICD9 (001-139) corresponds to infectious and parasitic diseases, and one of the subgroup is ICD9 (010-018), tuberculosis. Thus, ICD9 codes have a hierarchical nature. Some of the ICD9 codes are directly related to gender or patients of a certain age. For instance, ICD9 (630-679) are defined as complications of pregnancy and childbirth and ICD9 (600-608) stands for diseases of male genital organs. Thus, some demographic information can be extracted by looking at the ICD9 codes of the patients.

Data mining and analysis tools can be applied to EHR data to extract useful information such as computational phenotypes. However, most of the results from these computational tools will not be transformed to have practical clinical implications unless they can be inspected and validated by medical experts. It is therefore of utmost importance to design visual analytic tools to involve medical experts for interactive analysis of the extracted information, refinement of their analysis based on their domain expertise, and ultimately clinical decision making. In this paper, a sparse PCA based visual analytics tool is proposed for medical experts to analyze EHR data and allow them to interactively generate, interpret and refine computational phenotypes.

TABLE I: Notations used in this paper.

Notation	Explanation
d	Number of input dimensions
p	Number of output dimensions
$\mathbf{X}$	$d \times d$ covariance matrix
$\mathbf{Z}$	$d \times p$ orthogonal projection matrix
$\|.\|_F$	Frobenius norm
$\lambda_1(.)$	Largest eigenvalue of the input matrix
$\mathbf{z}$	Loading vector of the leading principal component
$\mathbf{w}$	A $d \times 1$ vector of normally distributed random numbers
γ	Regularization parameter
$\|.\|_1$	ℓ_1 norm

B. Phenotyping via Sparse PCA

Before we explain the details of the proposed phenotyping approach, we summarize the notations of this paper in Table I.

Given a patient cohort, identifying stable sub-populations, or phenotyping patients according to their health conditions characterized by their historial EHR (e.g., diagnoses and medication) can reveal important insights about the population. There are two prominent aspects of this process: 1) to explore deeper phenotypes at different levels of granularity, where phenotypes are naturally organized in a hierarchy, and 2) to closely and interactively involve medical experts in the generation, analysis, validation and refinement of the phenotyping process. Otherwise, the phenotypes may not be informative and thus lack the capability to generalize and provide further clinical guidance.

From the machine learning perspective, the problem of identifying patient sub-populations falls in the category of unsupervised learning since it does not depend on population related labels. To obtain patient populations, we can identify frequent clinical features and cluster patients accordingly. Since records in EHRs are encounters, some key clinical features such as diagnosis and medication can be identified based on their frequencies. However, we note that results from such univariate analysis do not consider complicated inter-dependencies among the features.

Principal component analysis (PCA) is a commonly used tool for unsupervised data analysis. Mathematically, finding the principal components is given by the following problem:

$$\max_{\mathbf{Z} \in \mathbb{R}^{d \times p}} \|\mathbf{XZ}\|_F^2, \quad \text{s.t. } \mathbf{Z}^T\mathbf{Z} = \mathbf{I}, \tag{1}$$

In PCA, columns of the matrix $\mathbf{Z}$ are dense. It means that the transformation with $\mathbf{Z}$ corresponds to the linear combination of all original features. These dense loading vectors of conventional PCA make it harder to interpret the output dimensions. When the purpose is dimensionality reduction and the features do not have specific meanings, this drawback is not critical. However, when the goal is to analyze and interpret the data by using PCA, it would be useful to know which of the original features contribute to the output dimensions. If the loading vectors are sparse, then only a few input features will be combined and this will make it easy to interpret the output dimensions. PCA with sparse loading vectors, that correspond to the columns of $\mathbf{Z}$ in Eq. (1), is called as Sparse PCA (SPCA) [11].

SPCA provides an integrated approach for phenotyping, for examples non-zero locations in the loading vector of the first principal component give a set of key clinical features for the

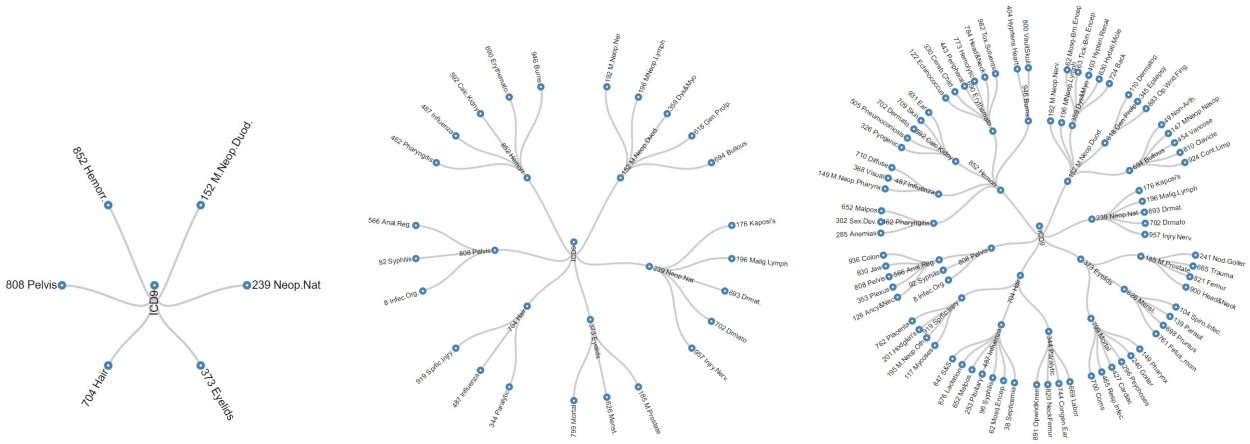


Fig. 2: An example of hierarchical phenotyping with stochastic Cvx-SPCA. The left most graph shows the first, the middle shows the second and the right most shows the third level features of the patient population. This procedure can be applied repeatedly until the desired number of levels is reached. When there are too many leaf phenotypes, users can manually inspect the properties of each phenotype and decide whether or not to expand it.

population. These key clinical features naturally define a set of subcohorts, each of which includes the patients associated with one of the features. Therefore, SPCA can be used as a relevant tool to obtain phenotypes and identify stable subcohorts, and its sparsity aids us to have a hierarchical representation which leads to a visualization approach. In the next section, we will show how SPCA can be used for phenotyping patients and how we visualize the obtained phenotypes in a way that physicians and medical researchers can easily interpret and refine the results.

C. Interactive Hierarchical Phenotyping via PHENOTREE

In this section, a phenotyping framework, PHENOTREE, is introduced to provide a visual analytic tool for analyzing, interpreting and refining computed hierarchical phenotypes. The proposed procedure for obtaining a two-level hierarchical structure using SPCA is explained in detail below:

Step 1: We apply SPCA to the whole patient population and obtain the non-zero loading values. Clinical features corresponding to the non-zero loading values are the input dimensions which contribute to the leading principal component. Therefore, these clinical features are selected as key features and a set of phenotypes within the population is defined starting with these key features. Thus, first level of the hierarchy is obtained using the key features determined in Step 1.

Step 2: First level features obtained in the previous step are used to define sub-populations containing the patients who have these key clinical features. Thus, the number of sub-populations in the second level equals to the number of first level features. In the second step, we use these sub-populations associated with each first level feature to expand their phenotype. Same procedure in Step 1 is applied to each sub-population and second level key clinical features are obtained for each first level feature.

By iteratively applying the method above to expand the phenotypes, we are able to build a phenotype hierarchy and organize it in a tree structure. Depending on how medical experts would like to explore the cohort, they can either 1) automatically grow all leaves of the tree, where the number of steps is identified by the size of the tree and the size of the tree can be determined by a physician, or 2) manually grow each phenotype leaf after investigating the properties associated with the phenotype (e.g., composition of the population, or according to a specific medical condition). More details about the interactive interface will be elaborated later. An important feature of the proposed phenotyping approach is the visualization ability of the hierarchical structure that medical researchers can easily observe the structures of the subcohorts, generate and refine phenotypes, and make clinical decisions. Otherwise, the aforementioned analysis is not possible with a text based representation of obtained phenotypes for a human since the number of patients and phenotypes are generally very large. For this purpose, PHENOTREE utilizes radial Reingold-Tilford tree [30] based on the work by J. Heer and J. Davies as given in Figure 2, where the outputs of steps 1, 2, and 3 for a three-level structure are shown.

Each node of this tree gives a structured phenotype and a stable subcohort characterized by this phenotype. We note that as we are expanding the phenotypes, the children are not independent from their parents because of the populations are conditioned on the parent phenotypes. For example, if the phenotype characterized by the diagnosis ICD9 92 Syphilis has a parent phenotype 808 Pelvis, we denote the phenotype as ICD9 $92 \rightarrow 808$. A patient may have a feature from only the first level, first two levels or from all three levels. For instance, in Figure 2, there are 3 patients who have ICD9 373 and ICD9 185, 32 patients who have ICD9 373 and ICD9 626, and one patient with ICD9 373, ICD9 185 and ICD9 761 features together. Same patients may have different hierarchical phenotypes as well. For example, one patient could simultaneously possess two phenotypes: ICD9 $373 \rightarrow 185 \rightarrow 761$ and ICD9 $373 \rightarrow 185$. If we need to assign patients exclusively to one of the phenotypes, the deepest hierarchy could be used. Thus, PHENOTREE provides an interesting, informative, and visually interactive way of phenotyping the patients by their diagnoses information. These

phenotypes can be used to cluster patients or can be used as side information for classification tasks.

The proposed approach to construct PHENOTREE is given in Algorithm 1, where the subroutine `PatientSampling` selects a sub-population in data with a specific phenotype, and `ExpandPhenotype` identifies a set of phenotypes of a finer level of granularity by solving the SPCA.

The proposed approach provides a novel way to investigate patient populations and to visualize the structure in a comprehensible way that medical researchers and physicians can observe the hidden information inside the EHR. Sparsity of the loading vector makes the interpretation possible because sub-populations are determined by the key clinical features which are the ones with non-zero loadings. Moreover, applying SPCA iteratively allows us to identify hierarchical phenotype structures and leads to an integrated way to visualize the structured phenotypes as well as different patient groups at various levels of granularity. Analyzing the various levels of granularity without visualization is a complicated and challenging task for a human expert.

As it was also discussed in [25], [27] and [28], cohort studies require interactive visualization approaches to provide insights of EHR datasets in a comprehensible way. Otherwise, medical researchers may face a risk of missing the significant information present in cohorts. As mentioned earlier, an interactive mode allows medical researchers to manually investigate and expand each phenotype. Therefore, an interactive interface has also been designed for hierarchical visualization shown in Figure 3. The purpose of this interactive interface is to provide a visual analytic tool for medical researchers so that they can generate the phenotypes, change parameters, interpret the results visually, and refine the findings. Refinement process has also an important role in medical research. In PHENOTREE, phenotypes do not have to be expanded uniformly. In the hierarchy in Figure 2, for example, not every key feature is expanded to the third level. In such cases, expertise of the medical researchers should be incorporated into the analysis process to expand the phenotypes further. Users of the interactive interface can upload an EHR dataset with features corresponding to the frequencies of the ICD9 diagnoses and generate a hierarchical visualization of the dataset by using this interface. Users can filter the records with respect to some properties such as demographic information of the dataset. Optimization technique and the SPCA formulation can also be selected according to the user's preference.

Algorithm 1 Construction of a PHENOTREE

Input: Data: $\mathcal{D}$, solver parameters for CvxSPCA: O, number of levels: N
Output: N-level PHENOTREE $\mathcal{T}$ and a set of phenotypes $\mathcal{P}$
1: Initialize tree $\mathcal{T} = \emptyset$
2: Add pseudo phenotype to phenotype stack $\mathcal{S}$: $p_0 \rightarrow \mathcal{S}$
3: **while** $\mathcal{S} \neq \emptyset$ **do**
4: Pop one phenotype p from stack $\mathcal{S}$.
5: **if** depth of p is less than N **then**
6: $\mathcal{S}$ = `PatientSampling` $(\mathcal{D}, p)$
7: Compute phenotypes of a finer level of granularity $\mathcal{P}_{(p,\mathcal{S})}$ = `ExpandPhenotype` $(p, \mathcal{S}; O)$
8: Update $\mathcal{T}$ with phenotypes $\mathcal{P}_{(p,\mathcal{S})}$
9: Push phenotypes in $\mathcal{P}_{(p,\mathcal{S})}$ to $\mathcal{S}$
10: **end if**
11: **end while**

With the proposed interactive interface, PHENOTREE can involve the medical researchers in the loop of computational phenotype. However, the interactive analytics and visualization constantly require performing deeper phenotyping on new identified stable sub-cohorts, and thus the phenotyping procedure will be invoked repeatedly. When applying the proposed method to analyze large- scale EHR cohorts, the efficiency of the SPCA will be critical. Unfortunately, existing formulations and optimization algorithms for SPCA are very sensitive to the scale of data, and the prohibitive computation time will cause huge delays when PHENOTREE is used interactively, and make the interactive phenotyping less practical. Therefore, we now propose an efficient SPCA method that is capable of exploring hierarchical phenotypes of large-scale EHRs.

D. Stochastic Convex Sparse PCA (Cvx-SPCA)

One concern about traditional SPCA methods is scalability with respect to huge amount of data points. Stochastic methods are generally preferred to deal with large sample sizes since only one gradient is calculated at each iteration. Therefore, we propose to use an efficient stochastic convex SPCA approach (Cvx-SPCA) studied in [31]. In this section, we summarize important points of Cvx-SPCA and refer readers to [31] for full details.

SPCA can be posed as an ℓ_1 norm regularized optimization problem as given in 2.

$$\min_{\mathbf{z}\in\mathbb{R}^d} -\mathbf{z}^T\mathbf{X}\mathbf{z} + \gamma \left\|\mathbf{z}\right\|_1, \tag{2}$$

A typical solution of such a problem with a composite smooth part and a non-smooth part is the proximal gradient descent such as [20], [21]. Even though methods such as [20] provide fast convergence properties, they compute the full gradient in each iteration. Therefore, the computation time increases drastically for large scale data with hundreds of thousands points such as EHRs. Therefore, stochastic proximal gradient descent (Prox-SGD), where only one gradient is computed at each iteration, is more suitable to deal with large scale datasets. However, stochastic methods generally suffer from low convergence compared to the full gradient methods. The reason of low convergence is the high variance due to random sampling. Therefore, a diminishing step size has to be used and that is why more iterations are required for the convergence of traditional stochastic methods. In literature, stochastic proximal gradient methods, which decrease the variance of the gradient computation and eventually lead to higher convergence rates, are also studied. For instance, proximal stochastic variance reduced gradient (Prox-SVRG) has been proposed to alleviate the low convergence downside of Prox-SGD in [24].

Prox-SVRG mitigates the effects of high variance by reducing it progressively and provides a geometric convergence rate. In Prox-SVRG, the important steps in gradient computation, which provide variance reduction, are using the average

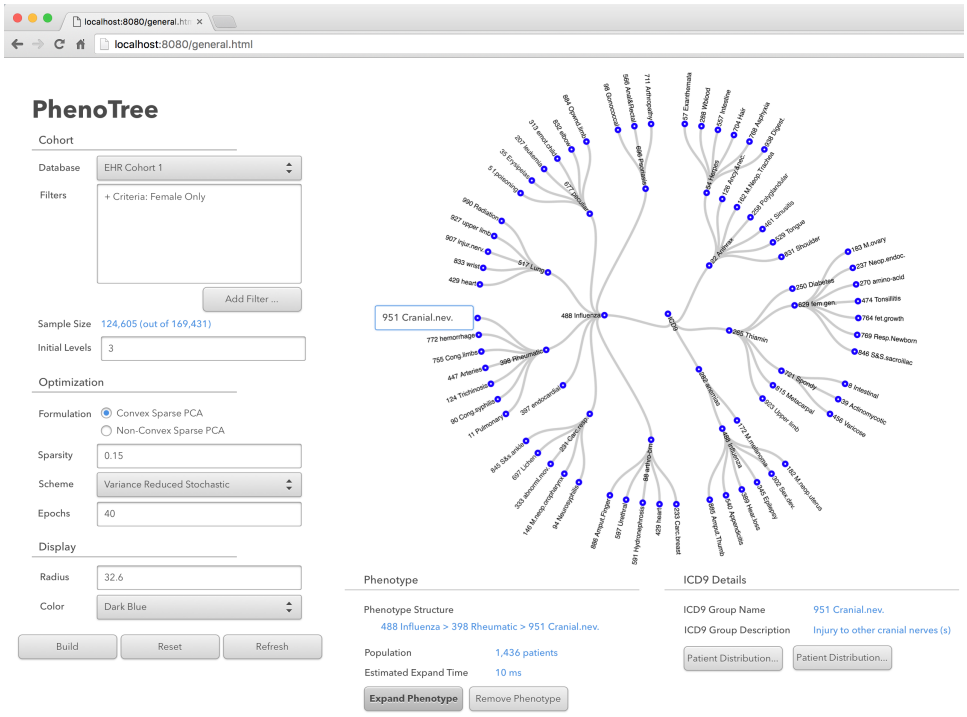


Fig. 3: An interactive interface for hierarchical visualization. User can upload a healthcare dataset whose features correspond to ICD9 diagnoses and generate a hierarchical visualization of the dataset by using this interface. Users can filter the records with respect to demographic information as needed. Optimization technique and sparse PCA formulation such as convex and non-convex can be selected.

gradient and incorporating the estimate of the optimum point. Estimate of the optimum point is the average of the points calculated during iterations and it is calculated at the end of each epoch to be used in the next epoch. Expectation of the gradient calculated in Prox-SVRG gives the full gradient which means that we are still in the direction of the full gradient under the expectation. However, the variance of the gradient is upper bounded since the variance is progressively reduced. As shown in [24], when the algorithm converges to the optimal point, variance of the gradient also converges to zero. Therefore, this approach can achieve better convergence rates than conventional stochastic gradient methods. We refer to [24] for detailed proof of bounding the variance.

Performance of Prox-SVRG is based on two main assumptions namely strong convexity and Lipschiz continuity of the gradient. Typically traditional PCA formulation is non-convex in nature as it can be seen from Eq. (1). A non-convex objective function is not suitable to leverage high convergence properties of Prox-SVRG. Therefore, we proposed to formulate the SPCA as the following convex optimization problem in [31]:

$$\min_{\mathbf{z}\in\mathbb{R}^d} \left(\frac{1}{2}\mathbf{z}^T \left(\lambda \mathbf{I} - \mathbf{X}\right)\mathbf{z} - \mathbf{w}^T\mathbf{z} \right) + \gamma \left\|\mathbf{z}\right\|_1, \qquad (3)$$

where $\lambda > \lambda_1(\mathbf{X})$ is the convexity parameter. Even though $\mathbf{w}$ is a random vector, we note that $\gamma = \left\|\mathbf{w}\right\|_\infty$ provides an upper bound of the minimal γ that gives trivial solution of an all-zero solution of $\mathbf{z}$, by subgradient analysis [32]. As such the sparsity tuning parameter can be normalized into a range of $[0, 1]$ by multiplying γ. This formulation without ℓ_1 norm regularization has been proposed in [19] as an approximation of learning the leading principal component. Eq. 3 provides a strongly convex formulation which transforms non-convex problem into a convex optimization problem. Thus, we can utilize Prox-SVRG for the solution of the optimization problem given above. One point we should note that it is usually hard to find strongly convex objective functions in many application domains. On the other hand, it is possible to show that the strong convexity assumption can be relaxed by using weaker conditions and a geometric convergence rate can still be achieved. For detailed convergence proof of stochastic Cvx-SPCA see [31].

In summary, we use an iterative stochastic gradient based framework to find the first principal component by learning a sparse loading vector. This approach is very convenient for large scale healthcare datasets, since it utilizes a stochastic framework. A comparison of the running times of our method with different SPCA algorithms is given in Figure 4. The efficiency of the proposed algorithm framework enables the interactive analytics and phenotyping via Cvx-SPCA. In the interface, we have shown users the approximate waiting time to expand a node in the phenotype tree (i.e., perform phenotyping based on the current population/sub-population). To achieve this, we use the data Table II to fit a linear regression model to predict the computation time (shown in Eq. (4)), which is then used to generate predictions.

$$\text{Time} = 10.95 + 0.0015 \times \text{Sample Size} \qquad (4)$$

For this purpose, several experiments with varying sample sizes are conducted and the execution times to construct a three

TABLE II: Running times (in seconds) for different sample sizes to construct a 3 level hierarchical visualization. Number of features affects the running time as well. This algorithm is not scalable in terms of feature dimensionality, since we need to calculate the covariance matrix. Therefore, running times shown here is for a fixed dimensionality of 500. A machine with 2.8GHz Intel(R) Xeon(R) CPU and 141.666 GB memory was used in the experiment.

Sample size	Running time (in sec.)
1,000	6.42
5,000	6.62
10,000	12.63
55,000	78.22

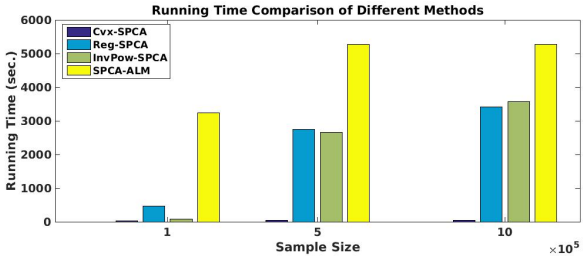


Fig. 4: Running times in seconds are compared to obtain similar cardinality of loading vector. A machine with 2.8GHz Intel(R) Xeon(R) CPU and 141.666 GB memory was used in the experiment. The method denoted by Reg-SPCA from [11], by InvPow-SPCA from [14] and SPCA- ALM from [15]. Reg-SPCA poses SPCA as an elastic net regression problem. InvPow-SPCA uses a generalization of the inverse power method with constraint optimization problems. On the other hand, SPCA-ALM uses semi-definite programming. According to experiments with randomly generated data with sample sizes of $100,000$, $500,000$ and $1,000,000$, we see that Cvx-SPCA can better handle large datasets than other methods.

level hierarchy are computed. Some example of sample sizes and their corresponding running times are given in Table II. Thus, we can estimate the running time of constructing a standard 3 level hierarchy by using the aforementioned regression model.

IV. Experiments

In this section, we will present the experiments conducted by using the proposed approach. The purpose of our experiments was to gain insight of the EHR dataset by illustrating the behavior of different patient populations. In the experiments, we applied the proposed visually interactive hierarchical phenotyping algorithm PHENOTREE to two real-world EHR datasets to build a hierarchical structure of the populations and visualize the obtained phenotypes.

A. Electronic Health Records Dataset

We used a private large scale EHR dataset comprising of $223,076$ patients and $11,982$ diagnoses. This EHR warehouse has records of patients over 4 years. Each diagnosis is represented by ICD9 codes. We do not have any explicit demographic information of the patients and we also do not know their admission/readmission times. Besides, some of the ICD9 codes have particular terms which give us a clue about gender and age of the patients. For instance, diagnoses about pregnancy, female/male genital organs, problems of newborns or diagnoses which have the term senile in their explanations can be used to group patients as female, male, child and old, respectively. However, it was observed that there are some patients who have records for both female and male specific diagnoses and some patients have diagnoses as being both newborn and senile. Since we do not have any control on the data collection phase, the main reason of this situation could not be resolved. Therefore, these kind of patients were eliminated in our experiments. There are also patients who have very few records in the dataset. Patients who have fewer than five records were also discarded. In the end, experiments were conducted with $168,431$ patients in total. Each patient has a sparse feature vector where the i-th value gives the frequency of the i-th diagnosis code for the corresponding patient. ICD9 codes correspond to different diagnoses and each diagnosis has its own sub-groups. For instance, 278 corresponds to Obesity and 278.01 is given for Morbid Obesity and 278.02 for Overweight, etc. Thus, sub-groups are all related to a main diagnosis. In our experiments, all the sub-groups related to a particular diagnosis were combined. As a result, the feature dimensionality was 927 in the experiments.

We conducted several experiments to visualize the structure of the patient population using the procedure explained in the previous section. A three-level PHENOTREE was generated as shown in Figure 6, where the ICD9 description of each disease is also included. In this visualization, it was observed that the following relationship can be established between layers. The first level diagnosis with ICD9 code 239 denotes Neoplasm Of Unspecified Nature. If we look at the output features of the patients who have ICD9 239, we can see ICD9 176 Karposi's Sarcoma, 196 Secondary and Unspecified Malignant Neoplasm of Lymph Nodes, 693 Dermatitis Due To Drugs, 702 Other Dermatoses, and ICD9 957 Injury to Other and Unspecified Nerves. Corresponding branches of the PHENOTREE can be seen in Figure 5. Karposi's Sarcoma is a type of cancer. Unfortunately, neoplasms or abnormal growth of tissue can spread out to different parts of the body. Therefore, patients who have diagnosis of neoplasm of unspecified nature may have other types of neoplasms as well. In addition, we can also see dermatological problems in the second level. Cancer treatments such as chemotherapy and radiation therapy can have side effects on skin and other organs. Especially, radiation dermatitis is one of the side effects of the radiation therapy. Another example can be the ICD9 344 Paralytic Syndromes whose second level were found as ICD9 669 Complications of Labor and Delivery, 744 Congenital Anomalies of Eye, Face and Neck, 820 Fracture of Neck of Femur and ICD9 891 Open Wound of Knee, Leg and Ankle. Paralytic conditions may not occur during delivery so much. However, methods like epidural may have the risk of paralysis. If we look at the second level features, we can also see that there are features related to neck or femur whose serious injuries can be a reason for paralysis. In the cohort, fractures and injuries were typical. Other most commonly encountered diagnoses were neoplasms

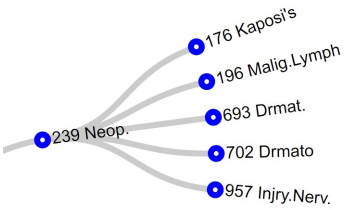


Fig. 5: A branch of the PHENOTREE. The first level diagnosis with ICD9 code 239 denotes Neoplasm Of Unspecified Nature. If we look at the output features of the patients who have ICD9 239, we can see ICD9 176 Karposi's Sarcoma, 196 Secondary and Unspecified Malignant Neoplasm of Lymph Nodes, 693 Dermatitis Due To Drugs, 702 Other Dermatoses, and ICD9 957 Injury to Other and Unspecified Nerves. Karposi's Sarcoma is a type of cancer. Patients who have diagnosis of neoplasm of unspecified nature may have other types of neoplasms as well. In addition, we can also see dermatological issues in the second level. Cancer treatments such as chemotherapy and radiation therapy can have side effects on skin and other organs.

(cancer), infectious diseases, and problems of newborns caused by complications of mothers.

After general patient population was analyzed, different patient groups in terms of age and gender were also investigated. Hierarchical representations of female, male, child, and old patient groups are given in Figures 7, and 8, respectively. Female and male groups were obtained by sampling the patients who have female and male diagnoses which are explicitly specified in ICD9 codes. However, we did not include all of the diagnoses that can be encountered among both females and males. Old and child patient groups were obtained in a similar manner. It is not possible to have an exact information about age of a patient from ICD9 codes. Some standard diseases which are mostly seen in children such as measles and problems of newborns are straightforward to classify as child. On the other hand, young patients, who are assumed to be adults younger than 70 years old in this paper, may have diseases such as hypertension which are mostly associated with people older than a certain age. Therefore, we picked patients with diagnoses that have the term senile in the ICD9 description as the old patients. As it is discussed, old and child patient groups are not defined by a certain age range. Child patients are assumed to be newborns and children before adolescence and the old patients are people older than 70. According to our observations, different patient groups tend to have the common diseases which was illustrated for general population earlier. On the other hand, different patient groups also yielded their specific diagnoses as well. For instance, one of the first level features in Figure 7a is ICD9 636, Illegal Abortion. Illegal abortion has many health risks for women such as infections and urinary tract disorders. In the second level, we can see diagnoses like ICD9 596, Disorders of Bladder, and 37, Tetanus, which could be side effects of illegal abortion.

Another example can be gleaned from the visualization of old patient population in Figure 8b. We observe diagnoses such as dislocation and fracture of bones. People older than a certain

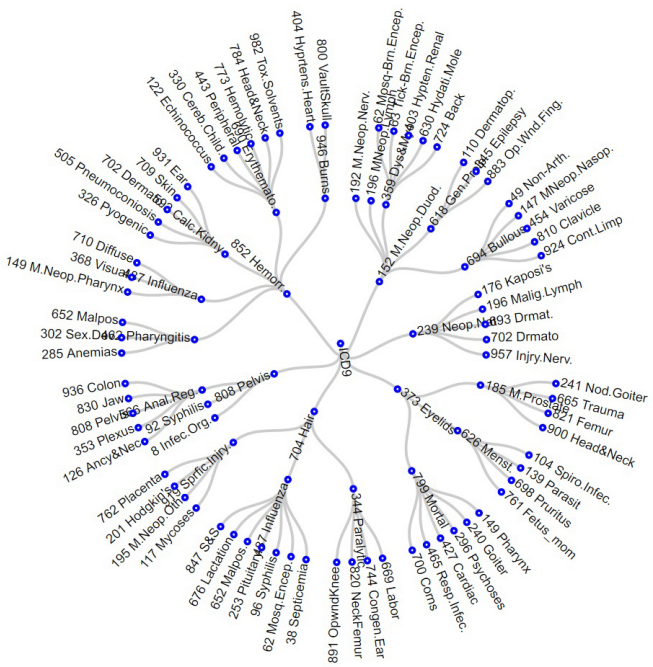
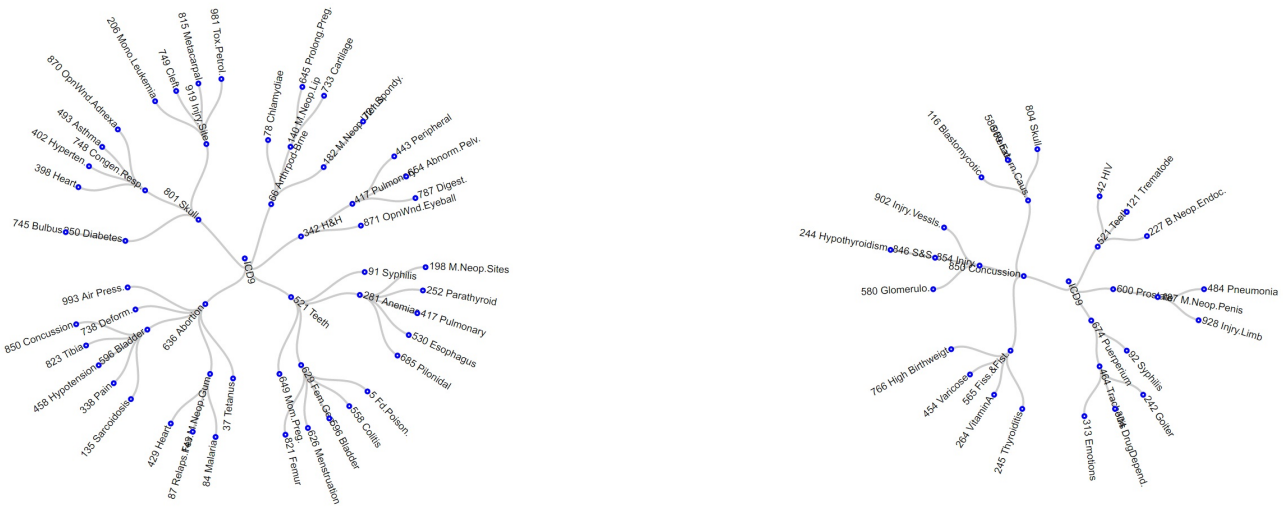


Fig. 6: Hierarchical stratification via Cvx-SPCA. Cvx-SPCA is applied on the entire patient population and the features with largest absolute loading values on the leading principal component are selected. Each feature dimension corresponds to a specific disease. A patient cohort is constructed by performing Cvx-SPCA repeatedly. Existing sparse PCA algorithms are unable to perform such analysis due to the scale of data.

age such as 80 commonly suffer from fractures especially in femur and pelvis. For example, ICD9 821 fracture in femur is one of the first level features of old patient group. In the second level of ICD9 821, diagnoses such as ICD9 268 Vitamin D deficiency, 332 Parkinson's Disease, some infectious diseases and ICD9 701 skin disorder were obtained. These diagnoses are commonly encountered among old patients. In the experiments with different groups of patients, not all of the diagnoses obtained were necessarily only about that specific group of patients. For instance, hierarchical stratification of female/male patients does not contain only female/male specific diagnoses. This is an expected outcome because we know that these patients may have records for other problems as well. As a summary, our results show that the proposed method can be used to understand and visualize the relationship between patient groups. PHENOTREE also provides a visual analysis of commonly encountered diagnoses and helps to understand the structure of the patient population for the EHR data.

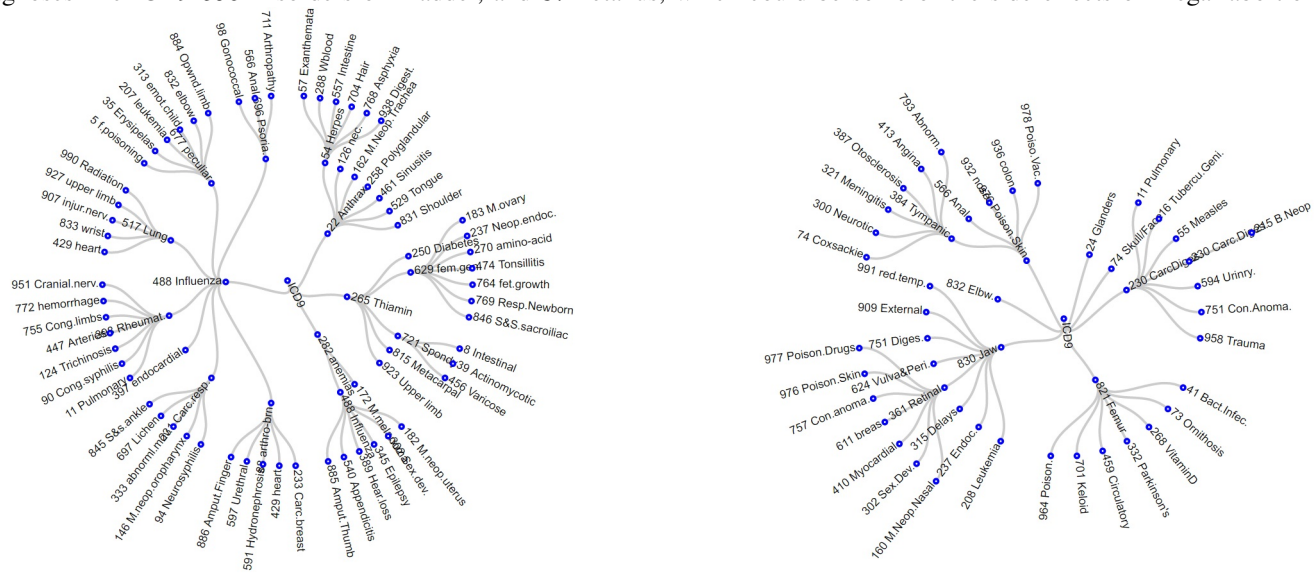
B. Diabetes Dataset

Next set of experiments were conducted for a public available EHR dataset which represents clinical data collected between (1999-2008) at 130 US hospitals [33]. There are $101,767$ patients in total and each patient has 50 features representing race, gender, age, number of medications, test results, diagnoses and other patient and hospital outcomes. Age distribution of the diabetes data is given in Figure 9. There is an age range for each patient, such as [0,10), [10,20), etc. In the figure, each age represents the corresponding range, for



(a) Female (b) Male

Fig. 7: Hierarchical stratification via Cvx-SPCA for female and male patients. Female groups were obtained by sampling the patients who have diagnoses pertinent to females which are explicitly specified in ICD9 codes. The same procedure was applied to obtain male patients. One of the first level features in Figure 7a is ICD9 636, Illegal Abortion. In the second level, we see diagnoses like ICD9 596 Disorders of Bladder, and 37 Tetanus, which could be some of the side effects of illegal abortion.



(a) Child (b) Old

Fig. 8: Hierarchical stratification via Cvx-SPCA for child and old patients. According to our observations, different patient groups tend to have common diseases which was illustrated for general population before. On the other hand, they also yielded their specific diagnoses as well. The age range is not defined precisely, since it is not possible to have an information about age from ICD9 codes.

instance, 10 indicates the range [0,10), 30 indicates the range [20,30), so on.

Each patient has 3 diagnoses with corresponding ICD9 codes. In the experiments, patients were represented by 729-dimensional feature vectors. First 697 features were corresponding to 697 unique ICD9 codes in the dataset. The value of the feature is 1, if the patient has that specific diagnosis. Therefore, the first 697 portion of the feature vectors is quite sparse. Rest of the features are provided with the dataset such as number of time in hospital, lab procedures, number of medications, and some test results about diabetes which are binary values. In addition, the readmission status of the patients is also given. Readmission attribute of the original dataset has three values such as not readmitted, < 30, and > 30 which correspond to readmission before and after 30 days of discharge [33]. We used this information to group patients as readmitted and not readmitted to see how the hierarchical structures differ.

We followed the same procedure as in electronic medical records dataset and we similarly aggregated the features belonging to the same ICD9 groups, resulting in 696 features in total. The results for the whole patient population are shown in Figure 10. Output features were observed as insulin and some ICD9 diagnoses such as neoplasm, heart disease, hormonal problem and so on. Insulin controls the blood sugar and therefore the existence of insulin indicates diabetes. If

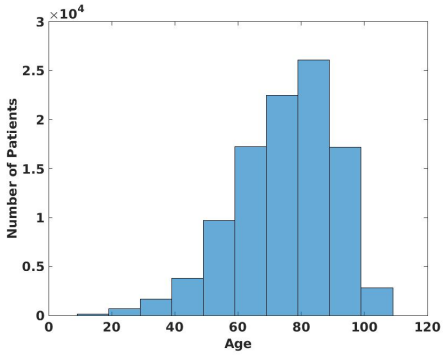


Fig. 9: Age distribution of diabetes dataset. We were not sure about the age distribution of the patients in the private dataset used previously. However, demographic information of diabetes dataset is available. Each age represents an age range, for instance, 10 indicates the range [0,10), 30 indicates the range [20,30), and so on.

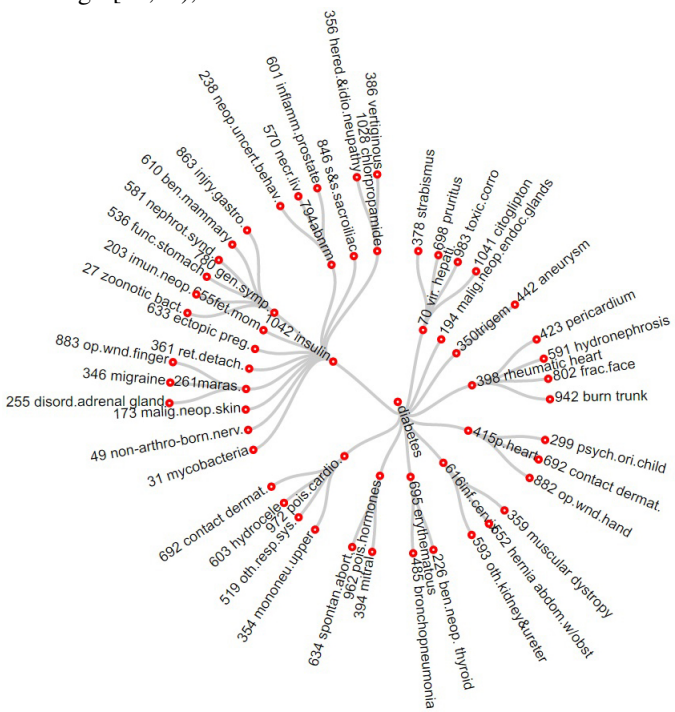


Fig. 10: Hierarchical representation of diabetes data. Proposed approached was applied to the whole patient population. Output features were observed as insulin and some ICD9 diagnoses such as neoplasm, heart disease, hormonal problem and so on. Existence of insulin indicates diabetes. Disorder of adrenal glands and stomach problems are commonly encountered among diabetes patients.

we further examine the diagnoses obtained from the patient groups who were prescribed insulin, we can see a wide range of diagnoses such as kidney problems, disorders of stomach, bacterial infections, and disorders of adrenal glands. Stomach problems are also commonly encountered among diabetes patients because of medications. Other than diabetes, it is possible to observe different kind of diagnoses as well. For instance, diagnoses such as 398 and 415 are related to the heart diseases which are frequently encountered problems.

Similarly, we examine the results of patients who are readmitted and not readmitted which are shown in Figure 11. Readmitted group has patients who were readmitted before and after 30 days of their discharge. The readmission is not only relevant for medical purposes but also for insurance companies [33]. From our point of view, we investigated how the types of diagnoses and the hierarchical structure of readmitted patients differ from the patients who were not readmitted. We should emphasize that the same regularization parameter was used for both patient populations. According to our observations, it is hard to distinguish these two populations by looking at the type of output diagnoses. For instance, diseases which may require the patient to get medical care regularly such as cancer are encountered in both results. However, if we look at the Figures 11a and 11b, we can see that the graph of readmitted patients has more nodes in the second level. The interpretation of this result can be made as follows: Readmitted patients may have several records for different diagnoses. Therefore, we could sample enough patients with specific diseases compared to not readmitted patient population, while we were constructing the levels. Graphs of female, male, old, teen and adult patients can also be seen in Figures 12 and 13, respectively. Patient groups with respect to gender and age were provided by the dataset.

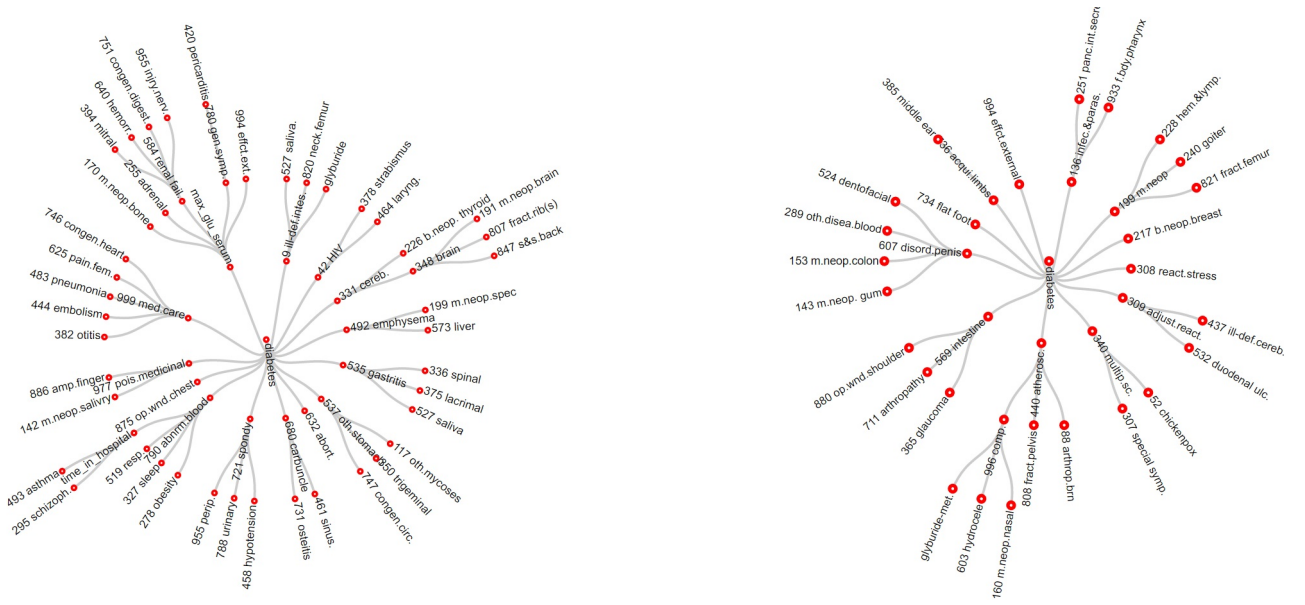
As a summary, we can see that exploring the insights and interpreting the findings about the EHR data visually are possible by using the proposed PHENOTREE approach. This kind of a system can be helpful for clinical decision support systems since it aids physicians to understand diagnoses and subcohort relationships in a visually interactive way.

V. CONCLUSION

In this paper, a hierarchical phenotyping approach for visualizing electronic health records is proposed by utilizing stochastic convex sparse PCA. The stochastic framework of the proposed approach enables applying PHENOTREE to large scale real world EHRs. Results show that proposed framework can be helpful to understand and analyze the patient populations and the relationship between them. Proposed approach also provides a way to visualize the relationships between patient groups with different diagnoses. In addition, it is shown that PHENOTREE can be used to detect clinical hierarchical phenotypes. We do not consider any temporal information of the records. In the future, we will incorporate the time stamps of the medical events in EHRs in our solution. When temporal information is considered, EHR data can be represented as multi-dimensional tensors, and we will extend SPCA to the tensor case and apply Prox-SVRG to obtain solutions efficiently. We will also design visualization for temporal phenotypes.

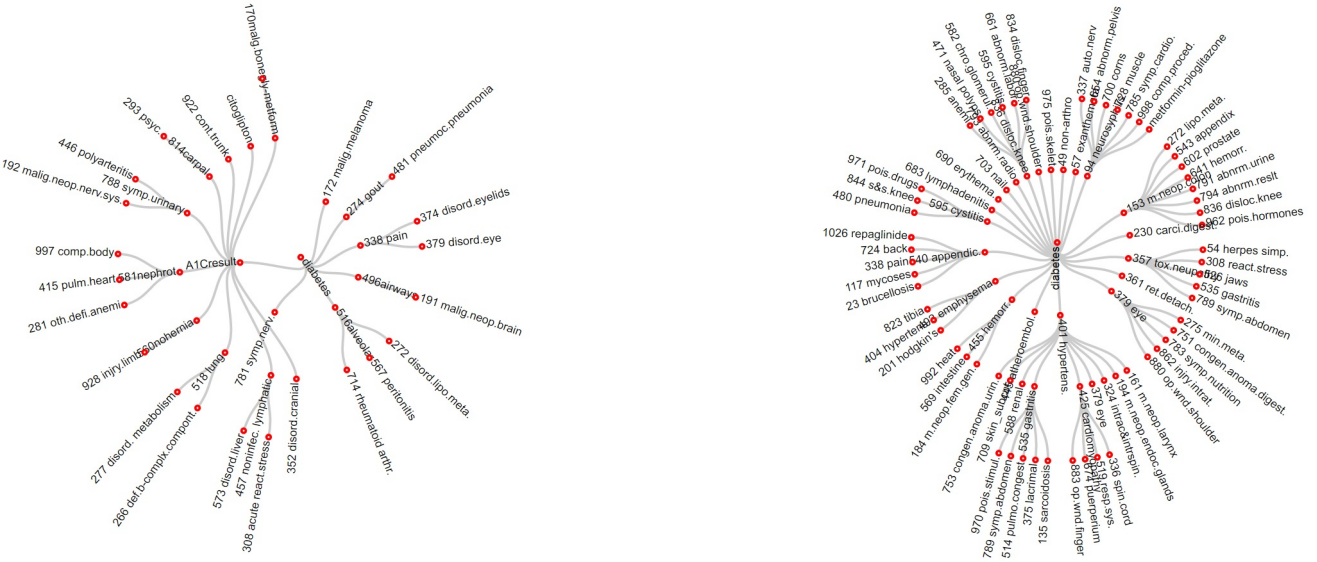
ACKNOWLEDGMENT

This research is supported in part by the National Science Foundation (NSF) under grants IIS-1565596 and III-1615597, and the Office of Naval Research (ONR) under grant number N00014-14-1-0631. The work of F. Wang is partially supported by National Science Foundation under Grant Number IIS-1650723.



(a) Readmitted (b) Not Readmitted

Fig. 11: Hierarchical Stratification via Cvx-SPCA for patients who were readmitted and not readmitted in diabetes data. Injury and poisoning are commonly encountered features for not readmitted patients. However, wider range of diagnoses such as neoplasms and diabetes melitus are observed for readmitted patients compared to not readmitted patients. In Figures 11a and 11b, we see that the graph of readmitted patients has more nodes in the second level, which may because that the readmitted patients have more diversed diagnoses.



(a) Female (b) Male

Fig. 12: Hierarchical Stratification via Cvx-SPCA for female and male patients in diabetes data. The information provided by the dataset was used to discriminate patients with respect to gender. We can observe female/male diagnoses along with common diseases. For instance, Figure 12b has ICD9 602 disorder of prostate and 401 hypertension. The number of nodes in Figures 12a and 12b are also different in this example.

REFERENCES

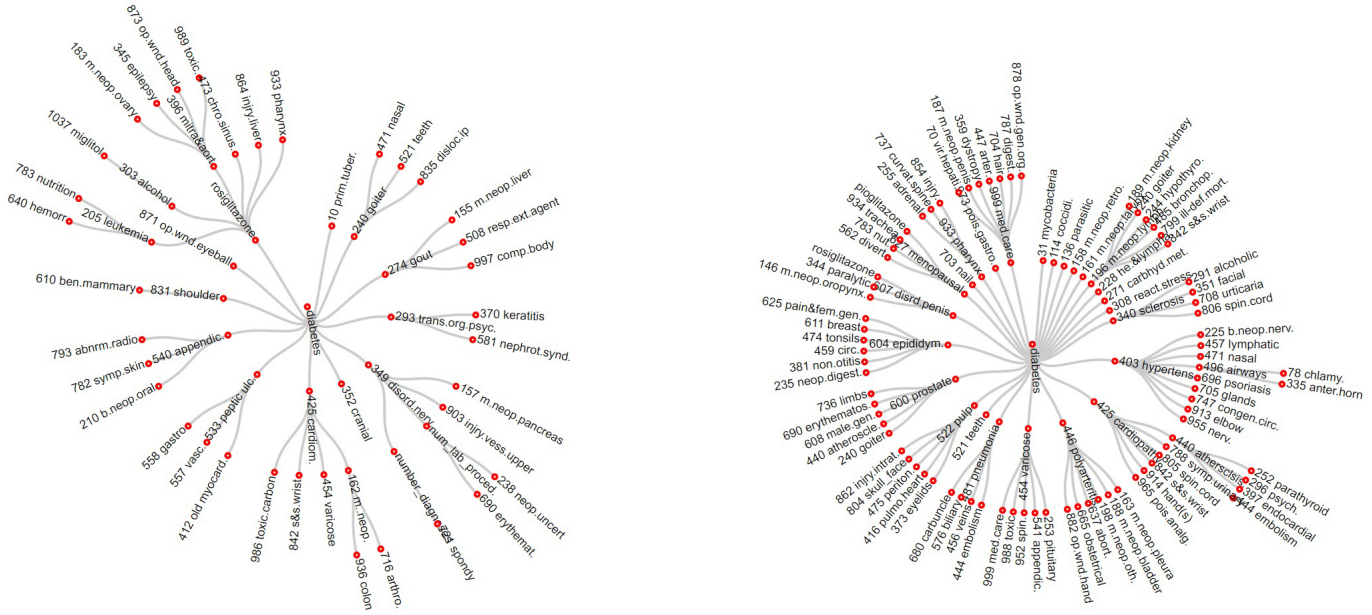
[1] C.-J. Hsiao, E. Hing *et al.*, *Use and Characteristics of Electronic Health Record Systems Among Office-Based Physician Practices, United States, 2001-2012*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012.

[2] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 135–144.

[3] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep computational phenotyping," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 507–516.

[4] J. C. Ho, J. Ghosh, and J. Sun, "Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 115–124.

[5] T. Xiang, D. Ray, T. Lohrenz, P. Dayan, and P. R. Montague, "Computational phenotyping of two-person interactions reveals differential neural



(a) Old (b) Teen-adult

Fig. 13: Hierarchical Stratification via Cvx-SPCA for old and teen/adult patients in diabetes data. Old and teen-adult groups of patients were determined by the age ranges provided in the dataset. Old patients comprises of the patients above 70 and teen-adult group consists of patients between 20 and 70. Neoplasms are very common in patient populations with a wide range of age or different genders for both datasets examined in this paper. We can see similar results in Figures 13a and 13b.

response to depth-of-thought," *PLoS Comput Biol*, vol. 8, no. 12, p. e1002841, 2012.

[6] [Online]. Available: https://datascience.nih.gov/bd2k

[7] R. P. Tracy, "deep phenotyping: Characterizing populations in the era of genomics and systems biology," *Current Opinion in Lipidology*, vol. 19, no. 2, pp. 151–157, 2008.

[8] Y. A. Lussier and Y. Liu, "Computational approaches to phenotyping: high-throughput phenomics," *Proceedings of the American Thoracic Society*, vol. 4, no. 1, pp. 18–25, 2007.

[9] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu, "Deep convolutional phenotyping," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. ACM, 2015, pp. 507–516.

[10] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 2012, pp. 389–398.

[11] T. H. Hui Zou and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[12] M. I. J. Alexandre d'Aspremont, Laurent El Ghaoui and G. R. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.

[13] e. a. Michel Journee, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, no. 3, pp. 517–553, 2010.

[14] M. Hein and T. Buhler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca," *Advances in Neural Information Processing Systems 23*, pp. 847–855, 2010.

[15] A. Y. Nikhil Naikal and S. S. Sastry, "Informative feature selection for object recognition via sparse pca," *IEEE International Conference on Computer Vision*, pp. 818–825, November 2011.

[16] Q. Gu, Z. Wang, and H. Liu, "Sparse pca with oracle property," in *Advances in Neural Information Processing Systems*, 2014, pp. 1529–1537.

[17] O. Shamir, "A stochastic pca and svd algorithm with an exponential convergence rate," *32nd International Conference on Machine Learning*, vol. 37, 2015.

[18] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[19] D. Garber and E. Hazan, "Fast and simple pca via convex optimization," *arXiv:1509.05647v4 [math.OC]*, November 2015.

[20] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Science*, vol. 2, no. 1, p. 183202, 2009.

[21] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[22] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," *Neural Information Processing Systems*, vol. 27, 2014.

[23] e. a. Nicolas Le Roux, "A stochastic gradient method with an exponential convergence rate for finite training sets," *26 th Annual Conference on Neural Information Processing Systems*, 2012.

[24] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.

[25] A. Perer, F. Wang, and J. Hu, "Mining and exploring care pathways from electronic medical records with visual analytics," *Journal of Biomedical Informatics*, vol. 56, pp. 369–378, 2015.

[26] D. Gotz, F. Wang, and A. Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of Biomedical Informatics*, vol. 48, p. 148159, Apr 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2014.01.007

[27] C.-F. Wang, J. Li, K.-L. Ma, C.-W. Huang, and Y.-C. Li, "A visual analysis approach to cohort study of electronic patient records," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2014, pp. 521–528.

[28] C.-W. Huang, R. Lu, U. Iqbal, S.-H. Lin, P. A. Nguyen, H.-C. Yang, C.-F. Wang, J. Li, K.-L. Ma, Y.-C. Li, and et al., "A richly interactive exploratory data analysis and visualization tool using electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, Nov 2015. [Online]. Available: http://dx.doi.org/10.1186/s12911-015-0218-7

[29] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: a review of the research literature," *International journal of medical informatics*, vol. 77, no. 5, pp. 291–304, 2008.

[30] Radial reingoldtilford tree. [Online]. Available: "http://bl.ocks.org/mbostock/4063550"

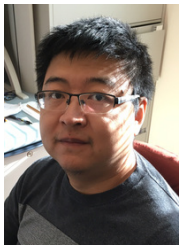
[31] I. M. Baytas, K. Lin, F. Wang, A. K. Jain, and J. Zhou, "Stochastic convex sparse principal component analysis," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2016, no. 15, 2016.

[32] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, pp. 2004–2005, 2003.

[33] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Research International*, 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Inci M. Baytas is a PhD student in the Department of Computer Science and Engineering at Michigan State University and a research assistant in Pattern Recognition and Image Processing Lab. She received her M.Sc. degree from the Electronics and Communication Department of Istanbul Technical University in 2014. Her main research interest is machine learning. Inci is currently working on distributed optimization techniques, biomedical informatics, multi_task learning, and distance metric learning.

Kaixiang Lin received his B.S. degree in The Department of Electronic Engineering and Information Science at University of Science and Technology of China in 2014. Since 2015, he has been working towards the Ph.D. degree in the Department of Computer Science and Engineering at Michigan State University. His research interests include machine learning and data mining, with applications to the large-scale gene and medical data.



Fei Wang is an Assistant Professor in the Division of Health Informatics, Department of Healthcare Policy and Research, Cornell University. His major research interest is data analytics and its applications in health informatics. He has published more than 150 papers in top data mining and medical informatics venues. He won the best student paper at ICDM 2015, best research paper nomination at ICDM 2010, Marco Romani Best paper nomination at AMIA TBI 2014, and his paper was selected into the best paper finalists in SDM 2011 and 2015. Dr. Wang is the vice chair of the KDD working group in AMIA.

Anil K. Jain is a University distinguished professor in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition and biometric authentication. He served as the editor-in- chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence (1991- 1994). He served as a member of the United States Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King Sun Fu awards. He is a Member of the National Academy of Engineering and a Foreign Fellow of the Indian National Academy of Engineering.

Jiayu Zhou is an Assistant Professor in the Department of Computer Science and Engineering at Michigan State University. He received his Ph.D. degree in computer science at Arizona State University in 2014. He has a broad research interest in large-scale machine learning and data mining, and biomedical informatics. He served as technical program committee members of premier conferences such as NIPS, ICML, and SIGKDD. His papers received the Best Student Paper Award in 2014 IEEE International Conference on Data Mining (ICDM) and the Best Student Paper Award at 2016 International Symposium on Biomedical Imaging (ISBI).