A Visual Analysis Approach to Cohort Study of Electronic Patient Records

Conference Paper · November 2014

DOI: 10.1109/BIBM.2014.6999214

4 authors, including:

Chih-Wei Grace Huang
Joint Commission of Taiwan
24 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Yu-Chuan Jack Li
Taipei Medical University
291 PUBLICATIONS **2,802** CITATIONS
SEE PROFILE

A Visual Analysis Approach to Cohort Study of Electronic Patient Records

Chun-Fu Wang Jianping Li Kwan-Liu Ma
Department of Computer Science
University of California at Davis

Chih-Wei Huang Yu-Chuan Li
College of Medial Science and Technology
Taipei Medical University

***Abstract*—The ability to analyze and assimilate Electronic Medical Records (EMR) has great value to physicians, clinical researchers, and medical policy makers. Current EMR systems do not provide adequate support for fully exploiting the data. The growing size, complexity, and accessibility of EMRs demand a new set of tools for extracting knowledge of interest from the data. This paper presents an interactive visual mining solution for cohort study of EMRs. The basis of our design is multidimensional, visual aggregation of the EMRs. The resulting visualizations can help uncover hidden structures in the data, compare different patient groups, determine critical factors to a particular disease, and help direct further analyses. We introduce and demonstrate our design with case studies using EMRs of 14,567 Chronic Kidney Disease (CKD) patients.**

I. Introduction

Electronic medical records (EMRs) provide massive amounts of patient data to physicians and clinical researchers, but there are no simple means for these medical professionals to fully exploit the data. Traditional EMR systems typically provide a tabular interface, which primarily models the underlying database rather than the high-level tasks that the user may desire to perform. For instance, when the user wants to identify the set of symptoms correlated to a chronic disease for a particular patient group, it requires manually browsing and searching through many records. Furthermore, mining EMRs has thus been increasingly raising attention for its potential to support patient stratification and uncover new disease correlations [1]. Clinical researchers can also use the data to verify existing knowledge and conduct exploratory study for new hypothesis. However, there is little support in current EMR systems for such clinical research tasks.

Visualization has been shown effective to present large amounts of information including EMRs [2]. Our overall goal is to create an interactive visualization system that can support a variety of EMR analysis tasks. We believe such a system can significantly raise the efficiency and performance of physicians and clinical researchers to make use of EMRs for their work. This paper presents our design of an interactive visual mining interface to cohort study of EMRs. This interface is a crucial part of the overall system to allow the user to make visual aggregation of the data for directing following analysis tasks. Our design addresses the high dimensionality [1], inhomogeneity [3], and complexity of EMR data. While we use a particular dataset in this paper to present our work, our interface supports a visual mining process applicable to EMRs in general and may be adopted by other analysis systems.

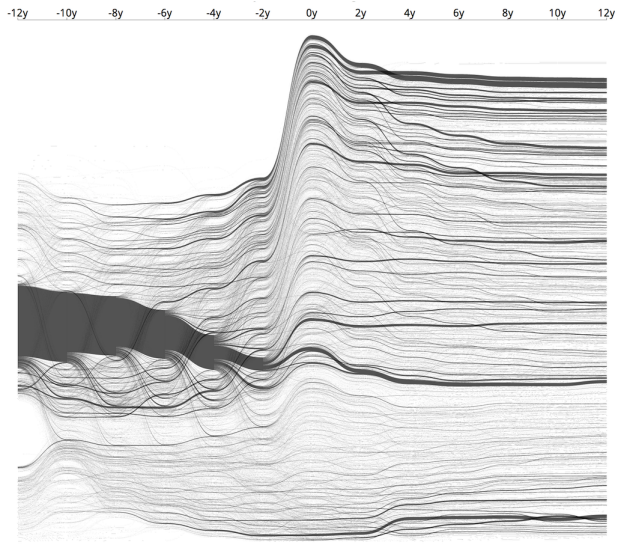


Fig. 1. Visualization of all 14,567 CKD patients clustered according to factor comorbidity. The X axis is time covering 12 years before and after each patient was diagnosed with CKD.

A. Driving Dataset

Disease co-occurrence (comorbidity) is of great interests in clinical research. With abundant EMRs, one can quantify comorbidities in a data-driven, statistical approach [1]. In addition, One can also understand the course of the disease by exploring trajectories of patient groups (cohorts), which are timelines consist of multidimensional, high variance variables derived from diseases, drugs or procedures. We call such variables the factors. Chronic Kidney Disease (CKD) is known for its correlation with a wide range of factors and the comorbidity varies over time stages.

The work presented in this paper has been driven by a data set obtained from Taiwan National Health Insurance Database (NHIDB), which contains ICD 9-CM (International Classification of Disease, Ninth Revision, Clinical Modification) codes for disease identification as well as the drug/procedure codes. From the one million population, we extracted 14,567 patients associated with CKD between year 1998 to 2011. Each piece of record stores the date, patient ID, and disease/drug/procedure codes.

For a visual-based cohort study of such a dataset, we are presented with several challenges. First, direct visualization of all the patients can easily lead to overplotting. Second, in this

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

dataset, there exist tens of thousands of factors pertinent to the CKD patients. It is not apparent how to discriminate and visualize these factors over time for bringing out structures of interest in the data. Fig. 1 shows a direct visualization of the observed factor comorbidity considering only 17 factors. As you can see, the visualization is already too complex to comprehend. It would be useful to select, aggregate, and visualize factors associated with patient groups. We have developed an interactive visualization system to support such operations.

II. Related Work

A. Temporal Visualization

Time information and timeline presentations [4] are traditionally of particular interest in analyzing EMRs. Many works have proposed presenting patient history with such longitudinal layout [5], [6], [7]. Real world data usually induce prohibitively high visual complexity due to high dimensionality or high variance. Thus, several simplification methods have been proposed. Bui et al. suggested using folder as well as non-linear spacing [8]. In the V-model project [9], Park et al. compressed causality relationship along the linear timescale to an ordinal representation to carry more contextual information of the event. In addition to abstracting time to use the horizontal screen real estate more efficiently, there are methods to save the vertical real estate. Bade et al. implemented a level-of-detail technique that presents data in five different forms based on its source and the row height available [10]. Our method simplifies the visual complexity of patient trajectories by aggregating records over time, clustering patients and filtering associations between cohorts.

B. Query-based Visual Analytics

In many real world cases, the user can narrow down the scope and reduce the complexity of the data by querying based on her domain knowledge. Systems of this kind allow the user to specify the pattern of interest [11] and can enhance the analysis process with advanced interfaces [12]. However, it is not always easy to translate an analysis task into proper queries [13].

For temporal event query, Wang et al. proposed an interactive system to support querying with higher level semantics such as precursor, co-occurring, and aftereffect events [14]. Their system outputs visual-oriented summary information to show the prevalence of the events as well as to allow comparison between multiple groups of events [15]. For overview-specific tasks, Wongsuphasawat et al. proposed LifeFlow, a novel visualization that simplifies and aggregates temporal event sequences into a tree-based visual summary [16]. Monroe et al. improved the usability of the system by integrating interval-based events [17] and developing a set of user-driven simplification techniques in conjunction with a metric for measuring visual complexity [18]. Wongsuphasawat et al. also extended LifeFlow into a Sankey diagram-based visualization, which reveals the alternative paths of events and helps the user understand the evolution of patient symptoms and other related factors [19].

In spite of their effectiveness in guided or well-informed analysis, query-based systems fall short for exploratory analysis where the user may not have a well-defined hypothesis and simply wants to explore and learn the data.

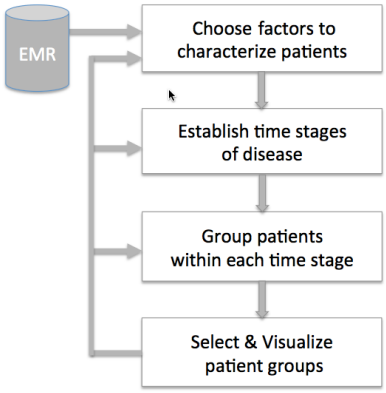


Fig. 2. Visual analysis process based on patient record aggregation.

C. Exploring inhomogeneous data

High-dimensional data items are less homogeneous comparing to each other. It is harder to associate/rank/filter those items meaningfully. Works had been proposed to slice-and-dice the data by dimension or item and separate them into homogeneous subsets [3]. It has been proven that, by carefully selecting projection methods, their system can incorporate multiple heterogeneous genetic data and identify meaningful clusters of patients [20]. Our work is a special case of the slice-and-dice concept, where we partition the record time into multiple stages (dimensions) and group patients (data items) within each time stage.

III. A Visual Analysis Process for Cohort Study

Our design starts by considering an analysis process to support cohort study of EMRs with respect to a particular disease. One challenge to address in this process is to discriminate the large and diverse comorbidity of factors. Lex et al. used the term inhomogeneity to describe such challenge and introduced a divide-and-conquer process to overcome it [3]. We follow this concept and design an iterative process to group patients such that the statistics of the factor comorbidity is relatively uniform within each group. The results give us cohort trajectories for the following study. Fig. 2 shows such an analysis process. It is an iterative process enabling the user to divide the data into homogeneous subsets that can be visually examined, compared, and refined.

A. Factors

The factors are derived from diseases/drugs/procedures and are the fundamental elements that characterize a patient in our system. In the CKD cohort study, there are tens of thousands disease/drug/procedure codes. Defining the right set of factors is not a trivial task because including unnecessary factors that are either redundant or irrelevant to the analysis objectives increases the computational cost as well as jeopardizes the legibility of the visualization.

Our system allows the user to define a set of factors by selecting independent codes or aggregating correlated ones based on her domain knowledge.

B. Time Stages

There are several known stages in the course of CKD, where the comorbidity structure remains relatively stable within a stage but varies from one stage to another. Such inhomogeneity between the stages could be the milestones of the cohort trajectories; in the meantime, the homogeneity within a stage helps clustering patient by comorbidity in the following analysis task.

Once the user specifies the time stages based on her domain knowledge, our system partitions the patient records accordingly. This is a human-assisted task because the homogeneity, especially on the semantic level, is often judged best by human [3]. A good initial guess would save unnecessary efforts; nevertheless, the user can always evaluate the results and refine the time stages. The results are patient trajectories regulated by time stages, which make the characteristics of each cohort easier to be analyzed.

C. Patient Groups

While the comorbidity of factors within each time stage is expected to be stable over time, its distribution over the population is not uniform. As a result, the population as a whole cannot be abstracted meaningfully and thus must be divided and analyzed separately.

Our system provides this capability with two clustering methods which aim at different aspects of extracting the underlying structures of the comorbidity distribution. The end results are cohorts of unique comorbidity. More detailed discussion is in Section IV-B.

D. Visual Examination

Once the time stages are defined and the cohorts are extracted, the quality of the abstraction can be visually evaluated by examining the associations between cohorts. For example, the user might want to examine the patterns how cohorts merge or diverge over time. Our system reveals the associations for the user to observe and interact with; however, the quantity or variance of the associations could be large and thus lead to visual clutter problems. Hence, our system also ranks and filters the associations based on their statistical importance, which is discussed in Section IV-B.

In any step of the visual analysis process, the user can go back and change the settings for factors, time stages, patient clustering and association filtering. For example, if the user wants to explore the temporal patterns in finer details and examine if there are local/short-term patterns, she can add more time stages to the context; on the other hand, if two or more stages exhibit indistinguishable patterns, she might want to merge those time stages as they do not convey extra messages. The user can also change the parameters to refine how patients are grouped or how associations are filtered. This iterative process continues until the user gets a satisfying result.

We use CKD cohort study to demonstrate the analysis process, but the process can be applied to the study of other diseases as well. For example, if the user wants to study the course of diabetes, she can define a list of factors related to diabetes. Then the user can apply the same process to set up time stages, cluster patients, and explore the structure of the cohort trajectories.

IV. System Design

The system design is based on the data transformation tailored to the analysis process. The transformed data are connected by a sequence of adjustable operators. We also discuss the rationale of the computational techniques applied. The visualization is driven by the transformed data, and the user feedback is directed to the corresponding operators and thus completes the iterative process. Finally, we discuss our visual and interaction designs to support the tasks.

A. Data Transformation

The data transformation steps behind the visual analysis process are illustrated in Fig. 3. The transformation order follows the analysis process from the raw patient records to the final visualization. Assume there are N patients and M unique factors. As the top-most chart shows, the raw sequence of a patient can be treated as a discrete timeline/trajectory with non-uniformly distributed records along the time axis. We define the patient trajectories as $\mathbf{P} = \{\mathbf{p}_1, ..., \mathbf{p}_n, ..., \mathbf{p}_N\}$ and the set of factors as $\mathbf{F} = \{f_1, ..., f_m, ..., f_M\}$. A patient trajectory is an ordered sequence of K_n records: $\mathbf{p}_n = (\mathbf{r}_{n,1}, ..., \mathbf{r}_{n,k}, ..., \mathbf{r}_{n,K_n})$, where each record consists of a factor set and a timestamp: $\mathbf{r}_{n,k} = (\mathbf{F}_{n,k}, t_{n,k})$, $\mathbf{F}_{n,k} \subset \mathbf{F}$. Note the timestamp of each record is not necessarily the record date. In the cohort study, we are interested in the temporal and populational patterns on the course of CKD. Hence, it makes more sense to align each patient trajectory by their days before/after being diagnosed with CKD.

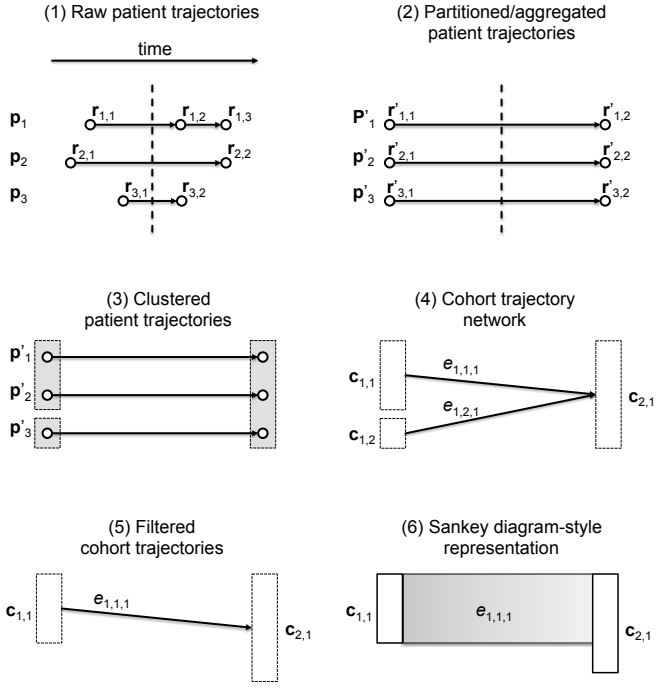


Fig. 3. Data transformation steps.

When the user specifies the time stages: $\mathbf{T} = (t_1, ..., t_l, ...t_L)$, the patient trajectories are partitioned based on their timestamps. Records in the same time stage are merged into one:

$$\begin{aligned} \mathbf{r}'_{n,l} &= \left(\mathbf{F}'_{n,l}, t_l\right) \\ \mathbf{F}'_{n,l} &= \bigcup_{i \in \mathbf{I}} \mathbf{F}_{n,i} \\ \mathbf{I} &= \{k | t_l \leq t_{n,k} < t_{l+1}\} \end{aligned} \tag{1}$$

The end results are patient trajectories regulated in time, $\mathbf{p}'_n = \left(\mathbf{r}'_{n,1}, ..., \mathbf{r}'_{n,l}, ..., \mathbf{r}'_{n,L_n}\right)$, where the timestamps are regulated by the time stages, and each record's factor set represents all the factors observed on that patient within the time stage. When the user requests for patient clustering, the patients at each time stage are clustered based on a certain similarity measure and become a set of cohorts: $\mathbf{C}_l = \{\mathbf{c}_{l,1}, ..., \mathbf{c}_{l,h}, ..., \mathbf{c}_{l,H_l}\}$, where $\mathbf{C}_l \subset \mathbf{P}$ and it represents a set of H_l cohorts at time stage t_l. We discuss the details of the clustering in the next section.

We define the cohort trajectory network as $G = (V, E)$, where each node $V = \{v_{l,h} | v_{l,h} = \mathbf{c}_{l,h}\}$ represents a cohort at a time stage, and each edge $E = \{e_{l,i,j} | v_{l,i} \rightarrow v_{l+1,j}, |\mathbf{c}_{l,i} \cap \mathbf{c}_{l,j}| > 0\}$ represents the association between two cohorts at consecutive time stages where their members overlap. The network G is used to drive the visualization in the end of the process.

B. Data & Control Flow

As shown in Fig. 4, data flow through a sequence of operators, which are adjustable and associated with different interactions by the user. The interaction workflow is designed from the user's point of view, and it implements the four tasks discussed in the analysis process in Section III: define factors, define time stages, group patients, and filter associations.

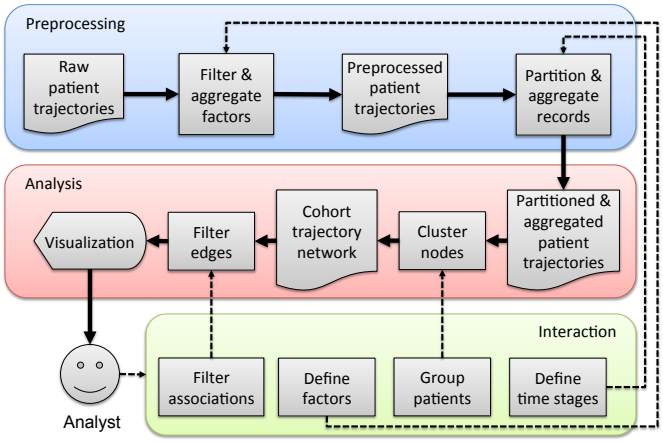


Fig. 4. The detailed data and control flow of the visual analysis process.

Once the user specifies the important factors for the study, the system scans the raw patient trajectories record by record and filters/aggregates the factors accordingly. Similarly, the time stages defined by the user also changes the way the system partitions and aggregates the trajectories by their time information. The two operators: *cluster nodes* and *filter edges* implement multiple techniques to support the analysis tasks of finding cohort and ranking/filtering associations, respectively. It is important to note that there is no once-and-for-all operation for any analysis task. Each cluster or filter operator has its strength and its limitations and thus should be carefully employed.

1) Frequency-based Cohort Clustering: Frequency-based clustering realizes the basic intuition of "See the main stream". Cohort with higher cardinalities are preserved while minor ones are considered less important and thus merged. Our system allows the user to specify a threshold x for the cardinality, and it merges cohorts of sizes less than the threshold into the "others" group.

$$\text{cluster}\left(\mathbf{C}_l\right) = \begin{cases} \mathbf{c}_{l,h} & \text{if } |\mathbf{c}_{l,h}| \geq x \\ \text{others} & \text{if } |\mathbf{c}_{l,h}| < x \end{cases} \tag{2}$$

2) Hierarchical Cohort Clustering: Given a time stage, each patient is characterized by the comorbidity of factors within the stage. We consider the similarity between two unique comorbidities as the set relation of their factors. For example, two sets of factors $\{f_1\}$ and $\{f_1, f_2\}$ are partially overlapped by the common factor f_1. In consideration of such similarity, we apply hierarchical clustering to extract cohorts of similar comorbidity.

The resulting clusters are hierarchical and the user has control to specify the desired number of clusters. With more clusters we're able to describe the characteristics of each cohort more accurately, but more clusters introduce more nodes, more associations, and thus higher visual complexity; on the other hand, fewer clusters create less visual complexity in the expense of potentially overlooking some essential structures.

Given the set of factors: $\mathbf{s}_i = \mathbf{F}'_{i,l}$ at a time stage t_l for a patient $\mathbf{p}_i$, we define the similarity between two patients with the Ochiai coefficient [21], which is a variation of cosine similarity between sets:

$$\text{similarity} = \frac{|\mathbf{s}_1 \cap \mathbf{s}_2|}{\sqrt{|\mathbf{s}_1|\,|\mathbf{s}_2|}} \tag{3}$$

3) Variance-based Association Filtering: The importance of an association lies on how confident we are able to make an inference from it. We can extract the statistically important associations by ranking and filtering their variances. Our system demonstrates this capability by adopting one particular type of variance, which is defined as the outcome entropy of the associated cohort. Such entropy can be calculated by the conditional probabilities of the different outcomes of the given cohort:

$$\begin{aligned} \text{pb}\left(\mathbf{p} \in \mathbf{c}_{l+1,j} | \mathbf{p} \in \mathbf{c}_{l,i}\right) &= \text{pb}\left(\mathbf{c}_{l+1,j} | \mathbf{c}_{l,i}\right) \\ &= \frac{|\mathbf{c}_{l,i} \cap \mathbf{c}_{l+1,j}|}{|\mathbf{c}_{l,i}|} \end{aligned} \tag{4}$$

$$\text{entropy}\left(e_{l,i,j}\right) = -\sum_k \left(\text{pb}\left(\mathbf{c}_{l+1,k} | \mathbf{c}_{l,i}\right) * \log_2 \text{pb}\left(\mathbf{c}_{l+1,k} | \mathbf{c}_{l,i}\right)\right) \tag{5}$$

We can see that the entropy is minimized when the patients in a cohort at the current time stage all go to another cohort at the next stage. In contrast, it is maximized when the

probabilities of patients go to other cohort are uniformly distributed. Our system allows filtering important associations by adjusting the entropy threshold. When the threshold is high, all associations are shown in spite of their variance; in the extreme case when the threshold is zero, only the associations of zero entropy will be displayed; in other words, it only visualizes the associations between fully overlapped cohorts.

C. Visualization Design

Our system visualizes the cohort trajectories network model we discussed in the previous section and presents it as an overview. The user can use it to assess important features such as cohort comorbidity, cohort distributions, and their associations across time stages, etc. We design the visual encoding and the optimization strategies in a way to maximize the legibility of the presentation.

1) Visual Encoding: We encode the dimensions of the visual space similarly to OutFlow [22], where the x-axis encodes the time information and the y-axis is used for laying out the categories (comorbidities). We also visualize the associations between the cohorts as ribbons.

The visualization must convey the characteristics of both the cohorts and the associations. It is common to encode cardinality to the node/edge [23], [24], as such information allows the user to assess the frequency-based distribution. Our system encodes cardinality as the node/edge height. Each cohort is labeled to show its dominant characteristics. It lists the common factors shared by all patients in this group. If there are factors not shared by the entire group, we indicate it by appending an asteroid to the label. In addition, we map colors to unique comorbidities and assign each node its corresponding color. The edge color is determined by the two nodes it connects, and we use gradient for smooth transitions.

The visual encoding of our system is tailored for the CKD cohort study; however, it can be easily changed to display other relevant information. For example, instead of showing the cardinality, the edge can encode other statistical measurements that reveal set relations [25].

2) Optimization: The overlaps between cohorts could be complex and thus increases the number of edges as well as the number of edge crossings. It could impact the legibility of the visualization [26]. Since the y-axis is nominal, and the ordering between the categories is flexible, we can arrange the node's vertical positions to reduce the number of crossing and thus resolve visual clutter.

The algorithm we apply to minimize edge crossing is modified from an existing library [27] and is a heuristic iterative relaxation method. The algorithm sweeps back and forth along the x-axis and adjusts the node vertical positions based on two objectives: (1) minimize the edge length, and (2) resolve node overlaps. It utilizes simulated annealing, so the process ends in a predictable time. The result is an approximation but the algorithm allows us to get reasonable results in an interactive rate.

In addition, the z-ordering(front to back of the screen) of the edges should be considered as well in order to maximize legibility [24]. We choose to place smaller edges on top of the larger ones to reveal the outliers.

TABLE I. FACTOR ASSOCIATION RULES

Disease (abbrev.)	ICD 9-CM/drug/procedure codes
Glomerulonephritis (GN)	582%, A350
Diabetes mellitus (DM)	250%, A181
Hypertension (HTN)	401%, A269
Hyperlipidemia	272%, A189
Polycystic kidney disease (PKD)	75312
Renal stone	5920, A352
Systemic lupus erythematosus (SLE)	7100, A431
Cerebrovascular disease (CVA)	430%-438%, A290-A294, A299
Coronary Artery Disease (CAD)	410%-414%
Congestive Heart Failure (CHF)	398.91, 402%, 404%, 425.4%-425.9%, 428%, A260
Chronic Kidney Disease (CKD)	585, 586, A350
Hemodialysis (HD)	58001C, 58019C, 58020C-58025C, 58027C, 58029C, 58030B
Peritoneal (PD)	58002C, 58009B, 58010B, 58011C, 58012B, 58017C, 58028C
Renal transplantation (RTPL)	V420
Proteinuria	7910, A469

D. Interaction Methods

The system interface consists of two views: trajectory view and summary view. The trajectory view displays the overview of the patient trajectories and the user can interact directly with it, such as selecting a group of patients. It also highlights the trajectories of the selected patients. Summary view presents the characteristics of the selected patient group. For example, it shows the distributions of gender, age, and factors, etc. It is also interactive and provides additional functions such as querying by patient meta information.

Most data items(patients, factors, etc.) in the system are selectable, and the system automatically searches for related items and highlights such associations with visual links. For example, the user can select a cluster of patients by clicking on a node or an edge in the trajectory view. The patients selected are highlighted as red regions in each node and link. The highlighted regions also encode the cardinality as heights so it shows the proportion of the patients selected comparing to others. In the meantime, the highlighted edges reveal the paths traveled by the selected patients. In addition, the user can also select a factor, and all patients having this factor will be highlighted. This enables the user to observe the global distribution of a particular factor.

V. CASE STUDIES

We work with nephrologists and define 17 factors as shown in Table I. Note that some of the factors are derived from the ICD 9-CM codes and others are derived from drug or procedure codes. The 17 factors represents the most related diseases and the important procedures that follow after CKD.

A. Explore Cohort Structures

In this study, we explore the global structures of the cohort comorbidity trajectories of the entire 14,567 patients data. We partition the records into multiple 2-year time stages. The analysts have different factor-of-interest for different stages of the course of CKD. In pre-CKD stage, they are interested in

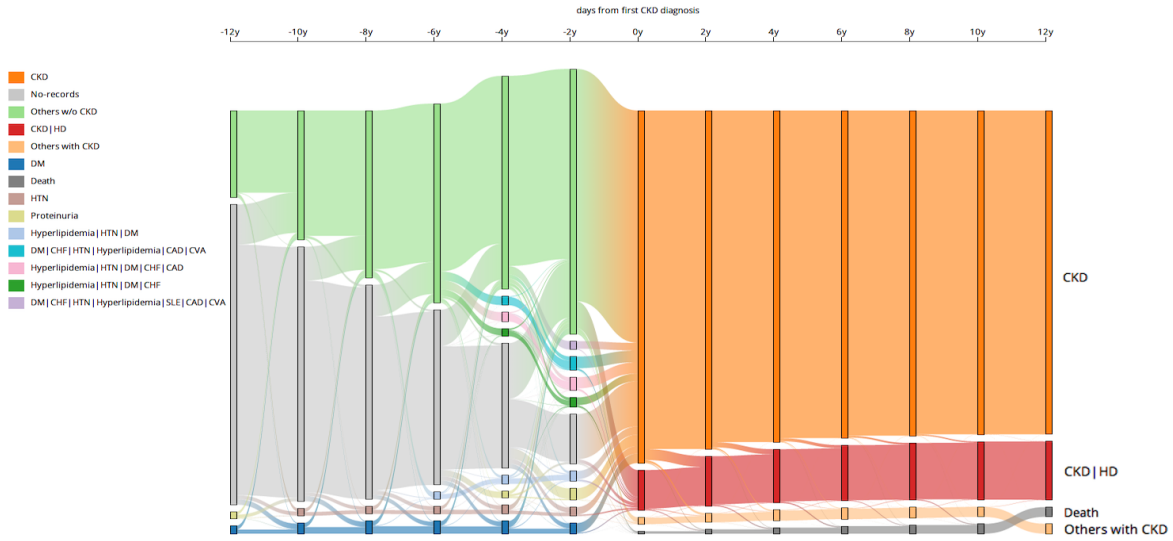


Fig. 5. Visualization of all 14,567 CKD patients clustered according to factor comorbidity. Smaller cohorts of size lower than 250 are aggregated.

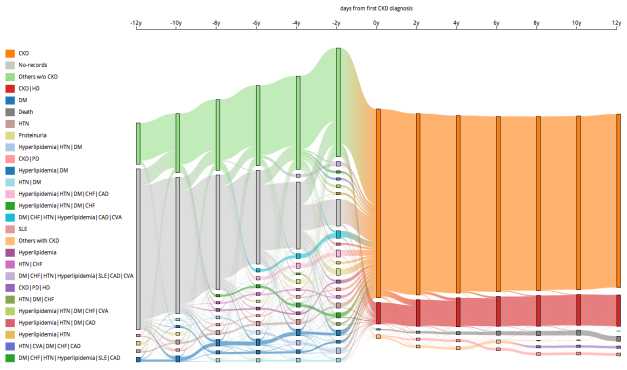


Fig. 6. Cohort trajectories with lower size threshold = 150.

common diseases such as hypertension, diabetes; for the post-CKD factors, they are interested in critical procedures such as dialysis, renal transplantation, or patient death. We filter the factors of each time stage as such. The patient records are pre-processed by a low-frequency filter to reduce unwanted changes over time; however, the detail of the filtering is outside the scope of this work.

Since there are too many unique comorbidities to visualize (as shown in Fig. 1), we apply frequency-based cohort clustering to extract the dominant cohorts. As Fig. 5 shows, the trajectories are simplified where larger cohorts are kept and smaller ones are merged into a single "others" group (light green for others without CKD and light orange for others with CKD).

From the overviews, we can learn about the prevalence of different comorbidities and their proportions in the population. For example, we can see from Fig. 5 that the cohorts of single disease such as hypertension (HTN)(brown) and diabetes (DM)(dark blue) shrink as the time close to year 0, which means patients start to exhibit other diseases. The user can lower the threshold to reveal smaller sized cohorts (Fig. 6).

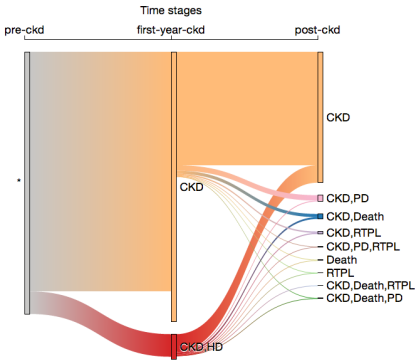
B. Explore Causal Relationship

In this study, we explore the causal relationship between hemodialysis (HD) in early stage of CKD and other factors. More specifically, we want to identify the driving factors that lead to hemodialysis and the possible factors caused by it.

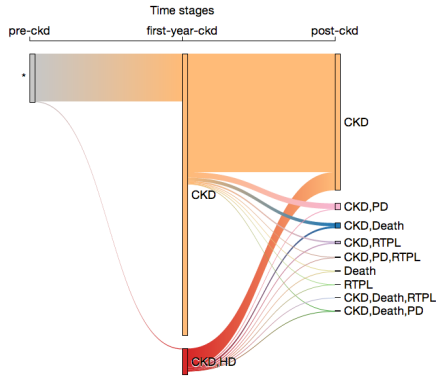
First, we consider three stages of the CKD disease: (1) pre-CKD: before the patient's first CKD diagnosis, (2) first-year-of-CKD, and (3) post-CKD: a year after of the patient's first CKD diagnosis. Second, we define the control factors for each stage based on existing knowledge. For the first-year-of-CKD stage, we focus on the factor: HD; for the post-CKD stage, we watch the other common factors of patient's CKD treatment: CKD, Death, PD, RTPL; for the pre-CKD stage, we watch all 17 factors. As a result, there are 835 unique factor combinations at the pre-CKD stage, two at the first-year-of-CKD stage and nine at the post-CKD stage.

Since there are only a total of 11 factor combinations at the first-year-of-CKD stage and the post-CKD stage, we can visualize it without any simplification process; however, there are too many combinations at the pre-CKD stage to be visualized directly. For simplicity, we first group them into one single cluster and focus on the last two time stages. As Fig. 7(a) shows, we find that 70.2% of the patients who took hemodialysis in the first year of CKD did not develop any other factors, while the rest of them either took peritoneal (PD) or renal transplantation (RTPL), or died. Some of the patients who were not taking hemodialysis in the first year also died; however, the mortality rate seems lower. We also notice that more than half of the patients who didn't take hemodialysis in the first year are not associated with any of the post-CKD factors. This means they were either in stable conditions (no prevalence of the factors we're watching) after the first year or their following treatments were not recorded.

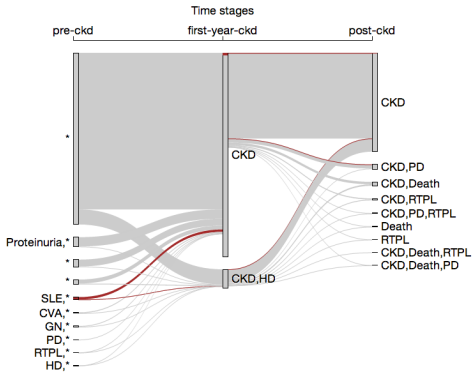
To see stronger causal relationship between the pre-CKD factors and HD at the first-year-of-CKD stage, we must filter out the associations that are not helpful. For example, if a group of similar patients are associated with both "CKD"



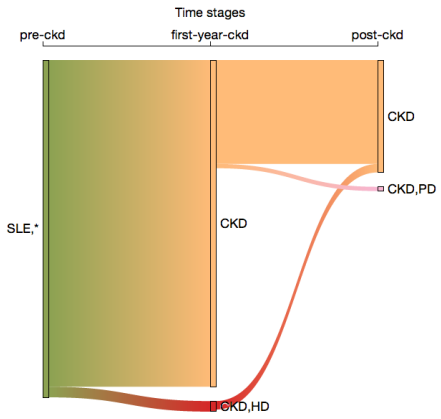
(a) Overview (Grouped pre-CKD into one cluster).



(b) Removed unconfident associations.



(c) Hierarchical clustering at pre-CKD stage (selected group "SLE,*").



(d) Focused on sub-group "SLE,*" (159 patients).

Fig. 7. Explore causal relationship (12,960 patients).

and "CKD,HD" clusters, it's harder to tell whether this combination of factors will cause hemodialysis or not. We can rule out all those unconfident associations by filtering the variance of their associations. We set a strict threshold 0.0 for the variance so that the association is kept only when it is 100% confident. After the filtering, 32.6% of the 835 unique combinations are taken out because their associations with the first-year-of-CKD stage are not confident. Fig. 7(b) shows that the remaining associations only covers 17.4% of the population. This means the pre-defined 17 factors might not be good explanatory variables to discriminate patients taking or not taking hemodialysis in the first year of CKD.

Next, we perform hierarchical clustering on the patients at the pre-CKD stage and generate ten groups of similar patients, as shown in Fig. 7(c). Note there are three groups labeled "*", which seems confusing at first as they could have been merged into one group. In fact, the three groups have different factor distributions. They are labeled "*" because none of the groups have a common factor shared by all members in the group. To avoid confusion, the user can assign custom label to describe the nature of the group. When we select and highlight the group who has a common factor of systemic lupus erythematosus (SLE), we find that none of them took the more serious procedures such as renal transplantation or died. Fig. 7(d) is a zoom-in view showing the structure of the selected "SLE,*" group. We also notice that the proportion of patients taking hemodialysis in the first year of CKD in the "SLE,*" group (3.14%) is one-third of such proportion in the entire population (9.54%).

C. Responsiveness

Our system is web-based and is tested with a commodity desktop machine (CPU: 2.66 GHz Quad-Core, Memory: 8GB 1066 MHz DDR3) as the application server and another desktop machine as the client. Most of the back end programs are written in Python, and the front end programs are written in Javascript and HTML5.

The system caches the transformed data after each operation in the data control flow (Fig. 4) to reduce unnecessary processing time and improve user end responsiveness. There are four major types of user interactions: defining factors, partitioning time stages, merging patients and filtering associations. The first two interactions usually happen at the beginning of a study and occasionally happen in major revisions; on the other hand, the later two types of interactions are much more frequent in the analysis process. Caching the less frequently updating results helps us reduce unnecessary processing time.

We measure the time elapsed for each process using the system timer. For 14,567 patients and 6,031,579 records, it takes 6 minutes to filter and aggregate factors of the entire data set, and 25 seconds to partition the data set into three time stages; however, such operations are taken only a few times throughout the analysis and thus do not require immediate response. More frequently performed operations such as clustering patients or filtering associations only take 5 seconds per time stage on average.

VI. Conclusions

We have developed a visual mining system to support explorative analysis of high-dimensional categorical EMRs. In

our CKD cohort study based on automated correlational analysis and human-assisted visual evaluation, our system allows the user to interactively assess the hidden structures of the cohort comorbidity trajectories. In addition, the analysis process is generalized and can be tailored for different disease studies and can work with different clustering/filtering algorithms.

As for the future works we would like to investigate the possibility of using more sophisticated feature extraction methods. In this work, we define the factors by hand with domain knowledge and group the patients based on the factors by a simple set similarity metric or a frequency-based metric. However, the combinations of factors are noisy and the variance within each cluster are usually high. Furthermore, there are still thousands of unused factors that may provide additional insights. Such problem could potentially be addressed with the help of correspondence analysis.

More can be studied for the visual encoding as well. Firstly, for conveying the association between the clusters, in this work we only visualize the cardinality of the association and filter them by variance. There are other measures of proportionality available [25], which can help evaluate the association from different aspects. We would like to study each method's role and effectiveness in conducting different analysis tasks. Secondly, for conveying and comparing the nature of each cluster, in this work we only present such information as text that shows the dominant factors of the cluster and indicates uncertainty; however, the underlying differences are non-binary and high-dimensional. How the system can effectively extract and present the subtle differences between the clusters could be the key to improve visual pattern detection.

Finally, it is possible to improve the computational performance by parallel data processing. As some of the steps in the analysis process are easily parallelizable while others, such as patient clustering, are not. We also intend to investigate more advanced database structures for efficiently data management.

ACKNOWLEDGMENT

This research is sponsored in part by the U.S. National Science Foundation and UC Davis RISE program.

REFERENCES

[1] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nat Rev Genet*, vol. 13, no. 6, pp. 395–405, Jun. 2012.

[2] B. Shneiderman, C. Plaisant, and B. W. Hesse, "Improving Healthcare with Interactive Visualization," *Computer*, vol. 46, no. 5, pp. 58–66, 2013.

[3] A. Lex, H.-J. Schulz, M. Streit, C. Partl, and D. Schmalstieg, "VisBricks: Multiform Visualization of Large, Inhomogeneous Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, 2011.

[4] E. R. Tufte and P. R. Graves-Morris, *The visual display of quantitative information*. Graphics press, 1983.

[5] S. B. Cousins and M. G. Kahn, "The Visual Display of Temporal Information," *Medical Informatics*, vol. 3, no. 6, pp. 341–357, 1991.

[6] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman, "LifeLines: using visualization to enhance navigation and analysis of patient records." *Proceedings of the AMIA Symposium*, vol. 08, no. 98, pp. 76–80, 1998.

[7] V. Nair, M. Kaduskar, P. Bhaskaran, S. Bhaumik, and H. Lee, "Preserving Narratives in Electronic Health Records," in *Proceedings of IEEE BIBM 2011*, pp. 418–421.

[8] A. A. T. Bui, D. R. Aberle, and H. Kangarloo, "TimeLine: Visualizing Integrated Patient Records," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 4, pp. 462–473, Jul. 2007.

[9] H. Park and J. Choi, "V-model: a new innovative model to chronologically visualize narrative clinical texts," in *Proceedings of ACM CHI*, 2012, pp. 453–462.

[10] R. Bade, S. Schlechtweg, and S. Miksch, "Connecting time-oriented data and information to a coherent interactive visualization," in *Proceedings of ACM CHI*, 2004, pp. 105–112.

[11] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: Timebox widgets for interactive exploration," *Information Visualization*, vol. 3, no. 1, pp. 1–18, 2004.

[12] J. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories," in *Proceedings of IEEE VAST*, 2006, pp. 167–174.

[13] J. Jin and P. Szekely, "Interactive querying of temporal data using a comic strip metaphor," in *Proceedings of IEEE VAST*, 2010, pp. 163–170.

[14] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, and S. Murphy, "Aligning Temporal Data by Sentinel Events : Discovering Patterns in Electronic Health Records," in *Proceedings of ACM CHI*, 2008, pp. 457–466.

[15] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith, "Temporal summaries: supporting temporal categorical searching, aggregation and comparison." *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1049–1056, 2009.

[16] K. Wongsuphasawat, J. A. G. G, C. Plaisant, and T. D. Wang, "LifeFlow : Visualizing an Overview of Event Sequences," in *Proceedings of ACM CHI*, 2011, pp. 1747–1756.

[17] M. Monroe, K. Wongsuphasawat, C. Plaisant, B. Shneiderman, J. Millstein, and S. Gold, "Exploring Point and Interval Event Patterns: Display Methods and Interactive Visual Query," University of Maryland, Tech. Rep. HCIL-2012-06, 2012.

[18] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, "Temporal event sequence simplification." *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2227–2236, 2013.

[19] K. Wongsuphasawat and D. Gotz, "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659–2668, Dec. 2012.

[20] C. Turkay, A. Lex, M. Streit, H. Pfister, and H. Hauser, "Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX," *IEEE Computer Graphics and Applications*, vol. 34, no. 2, pp. 38–47, 2014.

[21] A. Ochiai, "Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions," *Bull. Jpn. Soc. Sci. Fish*, vol. 22, no. 9, pp. 526–530, 1957.

[22] K. Wongsuphasawat and D. Gotz, "Outflow: Visualizing patient flow by symptoms and outcome," in *Proceedings of IEEE VisWeek Workshop on Visual Analysis in Healthcare*, 2011, pp. 25–28.

[23] R. Kosara, F. Bendix, and H. Hauser, "Parallel Sets: interactive exploration and visual analysis of categorical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 558–568, 2006.

[24] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive Sankey diagrams," in *Proceedings of InfoVis*, 2005, pp. 233 – 240.

[25] H. Piringer and M. Buchetics, "Exploring proportions: Comparative visualization of categorical data," in *Proceedings of IEEE VAST*, 2011, pp. 295–296.

[26] G. Ellis and A. Dix, "A Taxonomy of Clutter Reduction for Information Visualisation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1216–1223, 2007.

[27] M. Bostock, "d3 Sankey Diagram plugin." http://bost.ocks.org/mike/sankey/