1 INTRODUCTION

Electronic Medical Records (EMRs) provide rich data about individual patients and the opportunity to study a particular disease on a large population of patients. However, current EMR systems do not provide adequate support for fully exploiting the data. The growing size, high dimensionality, heterogeneity and complexity of EMRs demand a new set of tools for extracting information of interest from the data. Interactive visualization is one such tool [11] and this paper discusses some of the design considerations for building a visual analysis tool to study large EMR data.

A basic EMR system consists of two parts, namely the database (back end) and interface (front end). The interface allows physicians to access the database, and is typically graphical and based on multiple tabs with lists of information along with filters for refining searches. The interface closely models the underlying database, as opposed to the tasks that physicians must perform; that is, the interface is built with the data in mind, not the physician. For instance, physicians often want to know when a chronic disease was first diagnosed, which seems like a simple task, but in practice this requires manually searching through many electronic notes. There are many other examples of conceptually simple tasks that require extensive work on the part of the physician. Therefore, a key observation is that there is a large gap between the high-level tasks that physicians perform and the low-level database access that current EMR systems provide. This results in a relatively large workload for a physician reviewing patient records. The core problem is that physicians have very limited time to perform this extensive review. Therefore, designing a task-oriented process effectively supported by visual summarization of relevant data and an intuitive, easy-to-use interface is an essential topic.

Supporting a variety of interactive operations on very large and noisy temporal event data is challenging. Many existing visualization systems assume all the data can fit in memory, which is no longer feasible. Previous works to address the big data aspect of EMRs use data warehouse methodologies, where data from multiple EMR systems are first collected in a central relational database and then data marts are constructed according to analytics applications for fast access. Such a strategy, however, does not scale with the data size. Furthermore, existing EMR data review systems do not support interactive visualization tasks well. In fact, for visual analytics applications in general, considerations of data organization for efficient visual-based querying and interaction are lacking. How to manage and prepare EMR data to better support common analysis and visualization tasks is thus also an essential topic.

While previous research efforts have addressed many issues in visualizing patient records and temporal sequence data [10, 13, 3, 2, 8, 14, 7, 6], Lifelines2 [13] expedites the pattern finding process by aligning records by a sentinel event, where PatternFinder [9] increases the query capability by introducing chained events and time span. EventFlow [7] focuses on overview-specific tasks and defines a process to visually simplify the query summary; however, the above methods rely on a well-defined hypothesis to query and narrow down the scope. They are less suitable for exploratory analysis. The system we aim to develop should allow physicians or clinical researchers to make visual narratives of the EMRs in the process of data exploration. For example, we have been investigating how to employ hierarchical clustering methods and visual encoding strategies to bring out hidden structures for exploring causal relationships [12]. The resulting visualization will also help compare different patient groups, determine critical factors to a particular disease, and help direct further analyses. Furthermore, we place a focus on supporting cohort studies of EMR data and ensuring interactivity. In this short paper, we present our preliminary studies on cohort visualization and discuss data management solutions.

2 DRIVING DATA SET

Our work has used 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 patients, we extracted those associated with Chronic Kidney Disease (CKD) between the years of 1998 to 2011. There are a total of 14,567 such patients. Each piece of record stores the date and patient ID, along with the disease, drug and procedure codes.

With appropriate tools, this dataset will allow us to understand the course of a disease by exploring trajectories of patient groups (cohorts), which are timelines consisting of multidimensional, high variance variables derived from diseases, drugs or procedures. We call such variables the “factors”. CKD is known for its correlation with a wide range of factors and the comorbidity varies over time stages.

3 DESIGN CONSIDERATIONS

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, as shown by Figure 1. Second, in this 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. The visualization in Figure 1 provides an overview and could be used to guide further analysis, but it is too complex to be used by itself. It would be useful to select, aggregate, and visualize factors associated with patient groups. We discuss a few approaches to such operations and the design of an interface to match the intended tasks in [12]. To verify the usefulness of the resulting visualizations, we have conducted a few case studies followed by a small user study.
3.1 Case Studies

Figure 1 provides a broad overview of the data, but the visualization is hard to interpret. We simplified the visualization by always splitting patients into two cohorts. Note that the trajectory of each patient is aligned by the time when the patient was first diagnosed with CKD, which is marked as 0y. 2y means two years after the CKD diagnosis. A cohort of patients is represented as a vertical rectangular box with its height representing the cohort size and color indicating the different nature of the disease comorbidities in the cohort. A ribbon connecting one rectangle to another on its right represents the transition of patients from one cohort to another.

Figure 2 shows the results of simplifying the visualization by focusing on the groups of patients with and without hypertension (HTN) over time. We also mark those who died and those who had hemodialysis (HD) in black. We can see the number of patients getting HTN increased over time and became the majority two years before being diagnosed with CKD. The prevalence of HTN is irreversible, and it can be verified with the visualization because no patient moves from the cohort with HTN to the cohort without HTN. Finally, the patients with HTN seem to have a higher mortality rate and more likely needed hemodialysis.

Figure 3 shows visualization of those CKD patients with and without proteinuria. Again, the amount of patients receiving hemodialysis is marked in black. A larger increase in proteinuria patients can be observed over the 2-4 year period before being diagnosed with CKD. Receiving hemodialysis seems to be less dependent on having proteinuria or not. In Figure 4, we compare the use of two drugs: MA1 and A21. We can see patients started using A21 much earlier than using MA1. The medication did not seem to have particular impact on the patients.

We used these case study results to conduct a survey using nine subjects. Table 1 lists their job titles and areas of study/practice. One objective of the study was to find out if the subjects find the flow diagram visualization of the CKD data helpful or not. Almost all subjects find the visualizations in Figure 2 and Figure 3 helpful. For most of them, it's their first time seeing EMR data this way. They can immediately understand what each visualization shows. For example, they can tell from the visualization the correlation between medications and hemodialysis intervention. However, a few subjects had some difficulty interpreting visualizations in Figure 4. It turns out that they were confused by the crossing of the two groups, which is really a result of the layout. That is, this artificial pattern drew unnecessary attention from the subjects. We will therefore revise our design to avoid
3.2 Data Management Considerations

Interactivity is key to most of the visual analysis tasks, which often involve repeated filtering, aligning, aggregating, and rendering operations. A few promising approaches have been introduced to enhance interactivity of visual-based data querying and analysis. For example, Ermac [15] provides a declarative language for constructing interactive visualization, and imMens [5] relies on aggregation and pre-computation. EMR data, which have nested array and multi-value attributes, require special treatments. According to our experimental studies, conventional relational database cannot efficiently support all of these operations. We have adopted NoSQL [4] as the underlying database of our data management system because it offers more flexibility when we need to model different types of data such as physician’s notes, drug and procedure codes, image data, etc. NoSQL is also more scalable for handling large data because it can be used with distributed computing architectures.

We have been experimentally studying our data management strategies to support efficient processing of complex EMR data. We consider time as a unique characteristic of the data as what Aigner et al. have done in their conceptual framework for designing visualization systems for time-oriented data [1]. Unlike previous work which largely focuses on inter-record analysis [13, 6], we aim to better support inter-cohort analysis which demands advanced data management solutions to support fast data aggregation and reduction. The discussion of the specific design decisions for achieving required interactivity is beyond the scope of this paper.

4 Conclusion

There are many different ways to make use of the rapidly growing EMR data. Our study so far only addresses rather narrow aspects of the overall design problem for building a capable tool for studying large EMR data. However, our preliminary findings, the testbed system, and our continuing dialogue with the users of EMR data will keep our work in the right direction.

Our work promises a new means for physicians and medical specialists to more easily make sense of large amounts of data about their patients for making important decisions. The technology introduced will also make physicians rethink how they do their work, potentially leading to new directions and opportunities for optimizing their overall efficiency and productivity. We believe some of our designs will also be applicable to other temporal sequence data and heterogeneous data.

Acknowledgments

This research is sponsored in part by the U.S. National Science Foundation and UC Davis RISE Program. Thanks to Miss Chih-Wei Huang and Professor Yu-Chuan Li at Taipei Medical University for providing us with the CKD dataset and in assisting us to conduct the survey.

References


Table 1. Subjects in the user study.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Job Title</th>
<th>Area of Study/Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assistant Professor</td>
<td>Health Informatics</td>
</tr>
<tr>
<td>2</td>
<td>postdoc</td>
<td>Medical Informatics</td>
</tr>
<tr>
<td>3</td>
<td>Pharmacist</td>
<td>Pharmacy</td>
</tr>
<tr>
<td>4</td>
<td>Physician assistant</td>
<td>Internal Medicine</td>
</tr>
<tr>
<td>5</td>
<td>Project manager</td>
<td>Medical Informatics</td>
</tr>
<tr>
<td>6</td>
<td>Researcher</td>
<td>Health Informatics</td>
</tr>
<tr>
<td>7</td>
<td>Graduate student</td>
<td>Clinical Laboratory Science</td>
</tr>
<tr>
<td>8</td>
<td>Medical student</td>
<td>Medicine</td>
</tr>
<tr>
<td>9</td>
<td>Medical student</td>
<td>Biology</td>
</tr>
</tbody>
</table>
