Journal of Biomedical Informatics 56 (2015) 369–378

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Mining and exploring care pathways from electronic medical records with visual analytics

Adam Perer [a,*], Fei Wang [b], Jianying Hu [a]

[a] IBM T.J. Watson Research Center, 1101 Kitchawan Road, P.O. Box 218, Yorktown Heights, NY 10598, USA
[b] University of Connecticut, Storrs, CT, USA

ARTICLE INFO

Article history:
Received 3 December 2014
Revised 13 May 2015
Accepted 26 June 2015
Available online 2 July 2015

Keywords:
Visual analytics
Temporal event visualization
Frequent sequence mining
Data-driven care plans

ABSTRACT

Objective: In order to derive data-driven insights, we develop *Care Pathway Explorer*, a system that mines and visualizes a set of frequent event sequences from patient EMR data. The goal is to utilize historical EMR data to extract common sequences of medical events such as diagnoses and treatments, and investigate how these sequences correlate with patient outcome.
Materials and methods: The Care Pathway Explorer uses a frequent sequence mining algorithm adapted to handle the real-world properties of EMR data, including techniques for handling event concurrency, multiple levels-of-detail, temporal context, and outcome. The mined patterns are then visualized in an interactive user interface consisting of novel overview and flow visualizations.
Results: We use the proposed system to analyze the diagnoses and treatments of a cohort of hyperlipidemic patients with hypertension and diabetes pre-conditions, and demonstrate the clinical relevance of patterns mined from EMR data. The patterns that were identified corresponded to clinical and published knowledge, some of it unknown to the physician at the time of discovery.
Conclusion: Care Pathway Explorer, which combines frequent sequence mining techniques with advanced visualizations supports the integration of data-driven insights into care pathway discovery.

1. Introduction

Advances in data collection and storage technologies have led to ubiquitous yet complex Electronic Medical Records (EMRs). Because patient EMRs reflect the temporal nature of patient care, a patient's sequence of symptoms and diagnoses often correlates with their medications and procedures. These observed events may unlock the ability to analyze disease progression pathways and identify temporal patterns [33]. We believe such patterns may provide important insights into how diseases evolve over time and the effects of implemented interventions.

However, despite the fact that temporal knowledge discovery and pattern mining is no longer an unaddressed problem in data mining [1], it is still not easy to directly apply or adapt existing technologies to medical data for a number of reasons:

- There are many different event types encoded in EMRs. For example, there are thousands of distinct diagnosis codes, lab tests and drugs. Typically, large numbers of distinct event types can adversely affect the computational efficiency of temporal pattern mining techniques.
- EMRs may contain millions of patients over decades, and such voluminous data poses a great computational challenge to conventional methods.
- In medical scenarios, there are typically outcomes associated with each patient, such as the diagnosis of a disease or hospitalization. Clinicians are not only interested in the temporal patterns, but also in the correlations between such patterns and the patients' outcomes. Most existing pattern mining techniques lack the capability to elucidate such correlations.

In this paper, we propose *Care Pathway Explorer*, an interactive hierarchical information exploration system for analyzing longitudinal medical records. Our system provides a visual overview of frequent patterns mined from EMR patient traces. Instead of mining and visualizing all details at once, the interface supports interactive exploration for researchers to examine the level-of-detail relevant to user tasks by leveraging event hierarchies.

* Corresponding author. Tel.: +1 914 945 2681.
E-mail address: adam.perer@us.ibm.com (A. Perer).

http://dx.doi.org/10.1016/j.jbi.2015.06.020
1532-0464/

There are several major components to the *Care Pathway Explorer*:

- The Event Database, which stores patient electronic medical records with multiple event types, patient outcomes, and the event hierarchy for each event type.
- The Data Preprocessor, which constructs patient traces from the event database that can be directly fed into the Frequent Pattern Analytics engine.
- The Frequent Pattern Analytics, which mines frequent patterns from patient traces obtained by the Data Preprocessor and analyzes how these mined patterns correlate with outcomes.
- The Visual Interface, which provides visualizations of the frequent pathway events, with which users can interact to explore details of interest and generate insights.

These visual analytic technologies combine to support the mining and visualization of care pathway patterns. The goal is to provide insights into which practices lead to desirable patient outcomes, so clinicians can interpret meaningful patterns and customize care plans for complex patients. As established clinical practice guidelines typically only cover a single disease condition for average patients, such customization tools are critical in order to tailor care plans to the specific needs of real world patients who often have multiple complex comorbidities.

1.1. Background and significance

The exploration of temporal knowledge from longitudinal EMRs with data mining techniques is an important problem that has been the focus of study of much medical informatics research. In general, there are two types of studies: *holistic* and *localized*.

The goal of a holistic study is to exploit knowledge that can describe the overall event traces of the patient population. A typical technique that falls into this category is Business Process Management (BPM), which is a holistic management approach focused on aligning all aspects of an organization with the desires and needs of clients. In the healthcare domain, BPM technologies are mainly used for analyzing clinical pathways [2–4], which are standardized and normalized treatment patterns. However, applying BPM techniques to real patient data (e.g., for designing personalized clinical pathways) results in very complex and chaotic graphs.

To avoid the clutter caused by holistic studies, localized studies focus more on exploring the local characteristics of the patient event traces. For example, Norén et al. [5] propose a graphical statistical approach for summarizing and visualizing temporal associations between the prescription of a drug and the occurrence of a medical event, where the focus is the time period around the drug description. Chittaro and Combi [6] and Fails et al. [7] propose visual interfaces for constructing database queries to seek temporal patterns in multivariate temporal clinical data; the latter was further used in [8] for searching temporal patterns in patient histories. However, the system requires user specification of the structure of the pattern to constrain the database queries. Mörchen and Ultsch [9] propose a method called *Time Series Knowledge Mining* (TSKM) for uncovering local temporal relationships in multivariate data, but requires predefined temporal grammar and logic with prior knowledge.

Another methodology that is relevant to sequential pattern mining is *Temporal Abstraction* [10,11,35]. However, this technology generally requires an interval-based representation, which needs to know the duration of each event. In real-world EMR systems, duration information is often not captured, so we choose to use techniques that do not require this information. However, it is also possible to abstract point-based data by applying temporal knowledge which results in a more abstract representation of the data, in the form of symbolic time intervals Batal et al. provide several pattern mining techniques that uses a time interval-related representation of a sequence, which requires either the events have continuous values that can be quantized or the duration of every event is available [36,42]. Moskovitch et al. provide several approaches for discretizing continuous event values to derive more discriminative time-interval related patterns [40,41]. Patel et al. also provide a technique for mining interval-based events [43]. KNAVE II [12], VISITORS [13] and ViTA-Lab [34] are visual interfaces to interactively explore the temporal abstraction process in single and multiple patients, respectively. Other interval-based approaches include MuTIny [14] that discovers multi-time interval patterns, and MEMURY [15]. As most EMR data contains point events, and not interval events, our method aims to mine patterns from sequences of point events. Our work is different from these approaches in the sense that we only focus on point-based event sequences, although we propose a scheme for multiple levels-of-detail that could be applied to any type of pattern mining algorithm. We further note that pattern explosion can happen for either type of algorithm, whether it is for point-based event sequences or time-interval event sequences. Our scheme for multiple levels-of-detail, as well as our visual user interface, can be applied to these other type of pattern mining algorithms.

The *Care Pathway Explorer* system presented in this paper falls into the category of localized studies of EMRs. *Care Pathway Explorer* mines frequent patterns from patient traces and then illustrates them in a visually comprehensible and interactive user interface. There have recently been significant advances in the visualization community toward designing techniques for temporal event sequences of electronic health records. CareCruiser is a visualization system to compare EMR data to medical protocols [16]. LifeFlow [17] introduced a way to aggregate multiple event sequences into a tree, and EventFlow [18] later extended this approach to support both point-based and interval-based events. Outflow [19] designed a way to aggregate events into a graph, as well as integrating statistics. CoCo [37] is a tool for comparing event sequences at a cohort level.

Most recently, Frequence [20] is a user interface that integrates data mining and visualization in an interactive information exploration system for finding frequent patterns from longitudinal event sequences. The work described in this paper is an extension and adaptation of Frequence to support the use cases of medical informatics more directly, including a customized Event Database and Data Pre-processor designed for patient EMRs. Furthermore, an additional visualization was created to support an overall view of all events found in the patterns supporting a use case requested by physicians. In addition, Care Pathway Explorer has been integrated with a care plan template authoring tool, to support an end-to-end workflow from data-driven insights to institutional implementation.

2. Materials and methods

In this section, we introduce the *Care Pathway Explorer* system in detail, which supports the following flow of exploration:

1. The system shows an overview of the frequent patterns mined from patient event traces at the coarsest level, featuring statistics that indicate their correlations with outcomes.
2. The physician examines the frequent patterns and interactively selects specific patterns of interest for more detail.

3. The system computes the patient subsets that match the physician's specified sub-traces. The system then extracts the event traces for those patients, using a deeper level of the hierarchy, as specified by the user. These traces are then delivered to the frequent pattern miner engine.
4. The Frequent Pattern Analytics mines frequent patterns and displays them in the visualization alongside meaningful statistics.
5. This iterative process continues, returning to step 2 with each iteration, until physicians are satisfied with the insights generated.

Fig. 1 provides a graphical overview of the flow of the *Care Pathway Explorer* system, from which we can see that there are four major modules: *Event Database*, *Data Preprocessor*, *Frequent Pattern Analytics*, and a *Visual Interface*. In our implementation, the Event Database was implemented using DB2[1] as the relational database software. The Data Preprocessor and Frequent Pattern Analytics were implemented using custom analytics developed in Python. The web-based Visual Interface was implemented using the D3.js[2] visualization toolkit.

In the following sections, we introduce each of these modules in detail.

2.1. Event database

There are different types of events contained in patient EMRs which are stored in the Event Database. These events may be of many different classes, including labs, vital signs, medications and diagnoses. We integrate all classes of events in a single database using a *Universal Feature Model* (UFM), which is a four-column table that links patient IDs (unique for each patient), day IDs (timestamp of the event), event IDs (medical event) and an event value (numerical value associated with the medical event, if applicable). In the spirit of a star schema, there is also a separate patient table linking patient ID to the patient details, an event table linking event IDs to event details, and so on.

EMR data is processed and ingested using ETL (Extract, Transform, and Load) modules [21] to extract information from the EMRs and map them to coding standards and hierarchies, if available. Most medical events (e.g., diagnosis and medication) are organized hierarchically, which is leveraged in our system because we want to mine frequent patterns of medical events at different levels of detail to support iterative exploration. For example, Fig. 2 shows a set of events under cardiac disorders in the diagnosis hierarchy, which contains four different levels. The first level is the Hierarchy Name, which is the highest level in the Hierarchical Condition Categories (HCC) used in Medicare Risk Adjustment provided by Centers for Medicare and Medicaid Services (CMS). This level has 38 distinct event types. The second level is the more detailed Hierarchical Condition Categories (HCC), which contains 195 different codes. The third level contains 1230 unique Diagnosis (DX) group names (the first 3 digits of the ICD9 code). The fourth-level contains 14,313 different codes of the International Classification of Diagnosis 9th edition (ICD9). Similarly, medications can be mapped to a 3-level hierarchy using the United States Pharmacopeia (USP) Model Guidelines. The first level is the USP Category, which has 41 distinct high-level categories of medications. The second level is the USP Class, where has 129 distinct classes. Finally, the third level is the Drug Ingredient name, which has 5869 unique drug ingredients. All levels in the hierarchies are a many-to-one mapping to the higher levels.

[1] http://www-01.ibm.com/software/data/db2/.
[2] http://d3js.org.

2.2. The data preprocessor

The main goal of *Care Pathway Explorer* is to mine frequent temporal patterns from patient EMRs and explore them in a visual manner to reach insights. However, an issue that affects the efficiency of temporal pattern mining is when many events happen simultaneously. This is particularly true when the time granularity of the patient EMR is low resolution. Typically in EMRs, and especially in outpatient records, the finest time resolution is a day, and during a day, multiple medical events may occur to a patient. For example, it is rare that a patient undergoes only one lab test during a lab visit. Instead, a typical scenario involves a patient whose blood is drawn, with multiple lab tests conducted on the blood sample (known as lab test panels). Similarly, at a visit with a physician, the patient is required to have their height, weight, temperature and blood pressure measured, and this data is often inserted in EMRs as separate yet concurrent events.

Such data characteristics yield a great challenge for frequent pattern mining algorithms, as they detect patterns with all possible combinations of events and subsets of events occurring at the same time. We refer to this phenomenon as *pattern explosion*. To alleviate this problem, we preprocess patient traces before feeding them to the frequent pattern miner. The goal is to reduce the number of events happening at the same time. There are many *Same Day Concurrent Events* (SDCEs) contained in EMRs, thus we first detect the frequent *Clinical Event Packages* (CEPs) that are frequent subsets of SDCEs. If we treat each SDCE in every patient trace as a *transaction*, then the problem of detecting those CEPs is equivalent to the problem of frequent item-set mining [22], and each detected CEP can be used as a *super event*. Then, a greedy approach is applied based on *Two-Way Sorting* to break down each SDCE as a combination of regular and super events, such that the number of events contained in each SDCE is greatly reduced.

To better explain the process of breaking down SDCEs, we provide the following example: Suppose there exists a set of clinical events ABCDE that all appear on the same day. In order to avoid pattern explosion, this group of events needs to be broken down using commonly used CEPs that are detected. In this example the common CEPs are shown in the center of Fig. 3. The algorithm then sorts these packages according to the two-way sorting strategy. The CEPs are first sorted according to their cardinalities. Then, for packages with the same cardinality, they are sorted with respect to their appearance frequency. These sorting strategies represent the axes in Fig. 3. So, in order to break down ABCDE, the algorithm first finds the longest event packages that are subsets. In this case, ABC and ACE are the longest packages that are subsets of ABCDE. Then, because ABC occurs more frequently than ACE, ABC is selected as a super event contained in ABCDE. Besides ABC, the rest of the events are DE. Then the procedure is applied again to break down ABCDE as ABC, D, E. Using this technique, there are only 3 super events in ABCDE after the break-down procedure.

2.3. Frequent pattern analytics

After data pre-processing, the next step is to feed the processed patient traces to the *Frequent Pattern Analytics*. As the patient EMR traces can be very long, we choose *Sequential PAttern Mining with bitmap representation* (SPAM) [23] as our baseline approach. SPAM first constructs a bitmap style representation for all patient traces, so that all the event sequences are binary codes and the procedure of detecting frequent patterns involves *and*/*or* operations to these binary sequences. Due to this optimized strategy, SPAM has proven to be highly efficient for mining temporal patterns, especially from long sequences. That said, other frequent pattern analytics are also possible, such as Fradkin and Morchen's technique, which is a variant of the BIDE algorithm, for discriminative pattern

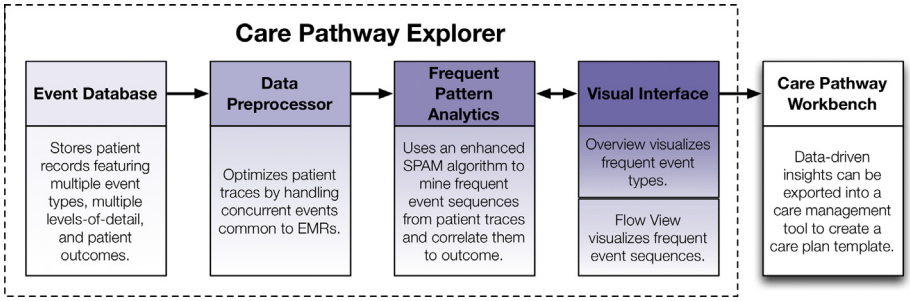

Fig. 1. An overview of the Care Pathway Explorer system, which contains four major modules: Event Database, Data Preprocessor, Frequent Pattern Analytics, and a Visual Interface.

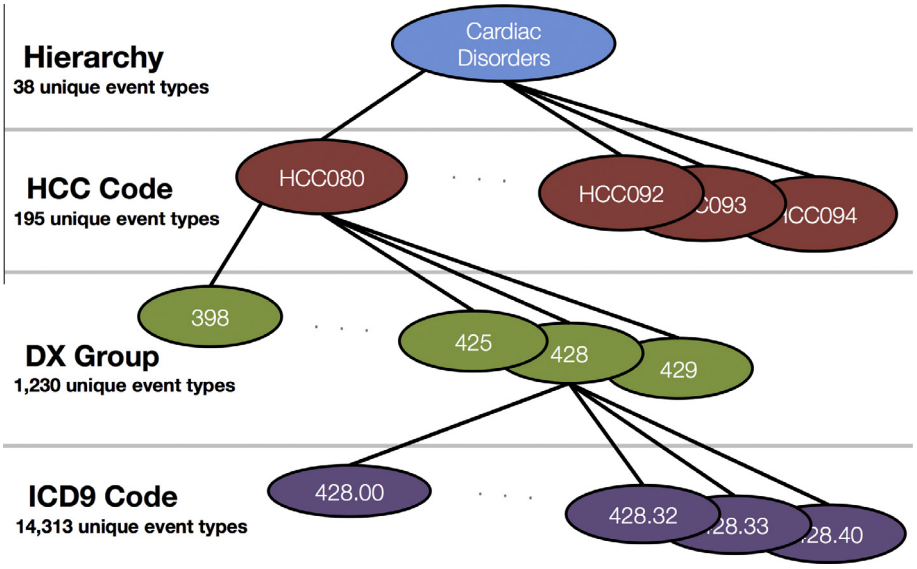

Fig. 2. An illustration of how many medical events (e.g., diagnosis and medication) are organized hierarchically in EMRs. This figure illustrates some of the events that and organized under Cardiac Disorders in the diagnosis hierarchy, which contains four different levels of detail.

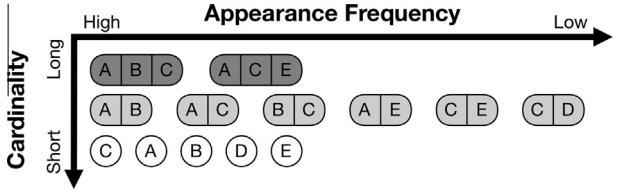

Fig. 3. An illustration of how the Data Preprocessor detects frequent Clinical Event Packages (CEPs) within Same Day Concurrent Events (SDCEs) to avoid pattern explosion.

mining [39]. Care Pathway Explorer's multiple levels-of-detail scheme for pattern mining can be applied to any specific approaches including BIDE.

However, in the medical domain, users may be interested in looking for patterns only within a certain domain-relevant time window. SPAM does not have the capability to incorporate such inter-event duration constraints. We thus modified the algorithm to provide this capability. SPAM also does not support outcome analysis directly, thus we extended SPAM to handle this capability in the following manner. Suppose each patient has an outcome, which can be either discrete (e.g., dead or alive) or continuous (e.g., HbA1c value for diabetes patients). In other words, every patient can be associated with an outcome of either positive or negative. If we have n patients, we can construct an n-dimensional vector, with the value on a specific dimension equal to 1 if the corresponding patient has a positive outcome, or equal to -1 if the corresponding patient has a negative outcome. For every pattern, we can also construct an n-dimensional vector with the value on the i-th dimension indicating the frequency this pattern appeared in the EMR sequence of the i-th patient. The percentage of the patient population that exhibits each pattern is referred to as the *support*, which indicates how frequent the pattern occurs. Then, we can compute the correlation statistics (e.g., Pearson correlation, odds ratio, relative risk and information gain) between every pattern vector and the label vector. That is how the correlation is computed. Such patterns are of particular interest to practitioners as they could represent potential best practices, or sub-optimal actions to be avoided, or sequences of events that indicate a particular risk. Additional details of modifications to the SPAM algorithm are described for Frequence [20].

2.4. Visual interface

The *Care Pathway Explorer* mines sequential knowledge from the EMRs so that physicians and clinical researchers can use these frequent patterns to understand disease evolution and optimize treatment plans. However, the quantity of patterns discovered is often very large. Thus, our system not only mines patterns but also presents the data in a user-centric way so that the patterns can be utilized in real-world settings. Information visualization is an effective way of communicating complex data, and thus *Care Pathway Explorer* features two complementary visualizations.

2.4.1. Overview visualization

After the pattern mining process, users can use the Overview Visualization, which resembles a bubble chart and displays events

of the most frequent patterns mined (Fig. 4 left). Each bubble represents a medical event that occurs frequently among the patients. The size of the bubble corresponds to the number of times each medical event occurs in the mined patterns. The color of the bubble also corresponds to outcome, so bluer bubbles are events that occur more often in positive patients (those who are not hospitalized with in the first year after diagnosis), whereas bubbles that are more red lead to patterns that are common with negative patients (those who are hospitalized within the first year after diagnosis). While the *X*, *Y* positions of the bubbles are abstract, each bubble is computationally positioned near event types with which it most frequently appears, to give an overview of clusters of patterns. The positioning is computed by first creating an event graph, where event nodes are connected by an edge if they occur in the same pattern. The edges are then weighted by the number of times the events appear in the same patterns. Within this graph, clusters of related events are identified by a greedy hierarchical clustering optimizing Newman's modularity metric [24]. These clusters are then positioned in the overview visualization using a hierarchical circle packing technique [25].

2.4.2. Flow visualization

In order to see how these bubbles connect to each other, there is a second visualization that resembles a Sankey diagram [26], a type of chart often used to show the magnitude of flow between nodes in a network. We describe the characteristics of our visualization using Fig. 5 as an illustrative example.

Events in the frequent patterns are represented as nodes, and event nodes that belong to the same pattern are connected by edges. For instance, the simple pattern **Diagnosis → Medication**, is visualized as a **Diagnosis** node connected to a **Medication** node at the bottom of Fig. 5. Patterns that share similar subsequences, such as **Lab → Diagnosis → Medication** and **Lab → Diagnosis → Lab**, involve two edges from **Lab** to **Diagnosis** representing each subsequence. Thus, prominent subsequence patterns also become visually prominent due to the thickness of the combined multiple edges. The Sankey-style layout is computed using iterative relaxation. First, the horizontal position of each node is fixed based upon its position in the pattern. Next, the layout algorithm begins with the nodes on the far left, and then places the connecting nodes on the right to minimize edge distance. Then, the iterative process has a reverse pass, going from right-to-left, and then the entire process is repeated five times. Node occlusion is avoided by shifting nodes that overlap due the result of the layout algorithm.

Of course, not all patterns are equal, as some correlate to good outcomes whereas others correlate to bad outcomes. This visualization uses the same color mapping as the Overview visualization. By default, the outcome is determined by Pearson correlation, but users can interactively select other measures (e.g., odds ratio, relative risk and information gain). Users can also mouse-over edges to get additional data, including a description of the pattern and statistics describing the patients.

2.5. Interaction

The visualization is organized hierarchically, based on the Event Database. Initially, the visual interface is populated with all frequent patterns at the coarsest level. The overview visualization acts as a starting point for users to interact with the visualization and explore patterns of interest. Users can click a sequence of nodes or edges to highlight an interesting pattern. This selection enables a query for all patients who have traces that fit the pattern. Users can explore the patterns of all patients, or explore their patterns in more detail, by drilling-down to the next level of hierarchy to get more specific information. For instance, if users select the pattern **Diagnosis → Medication**, the visualization shows all patients that matched the pattern, and then the mined pathways would be visualized in more detail using diagnosis HCC codes and medication Pharmacy Subclasses. The user can make selections and hierarchically drill down until the desired level-of-detail is reached. If a user would like to focus on patterns of a particular support or outcome range, users can use the range sliders to filter to the patterns of interest.

After extended analysis, if a care plan coordinator determines a set of patterns would be a beneficial sequence to add to the healthcare institution's care plan template, the user can select the pattern and deliver it to a care plan template tool. In order to demonstrate this concept, we have integrated *Care Pathway Explorer* into a care plan template tool, Care Pathway Workbench [27].

3. Results

In order to evaluate the utility of *Care Pathway Explorer*, we conducted a long-term case study of a real physician using real-world datasets to demonstrate its effectiveness at reaching insights in practice. Research in the visualization community suggests that traditional evaluation metrics (e.g. measuring errors or task time completion) are often insufficient to evaluate visual analytics systems designed for data exploration [28–30]. Instead, we chose to use the evaluation methodology developed by Perer and Shneiderman [31] and conducted a long-term case study with Dr. Robert Sorrentino, the Chief Medical Officer of Providence Medical Foundations.

Dr. Sorrentino was interested in analyzing patterns mined from patients with hyperlipidemia. Hyperlipidemia is a common chronic condition and several large clinical trials have identified LDL (Low Density Lipoprotein) as one of the major predictors of heart disease. While there are various medications for lowering LDL levels, comprehensive guidelines on the most effective intervention are still lacking, particularly for patients with multiple comorbidities. Analyzing treatment patterns and associated outcomes of such patients could provide valuable insight into how to customize the general guidelines to achieve better outcomes. Dr. Sorrentino was thus interested in analyzing the patterns consisting of all medication events, as well as diagnosis events related to hyperlipidemia. The groups of the diagnosis codes mined with *Care Pathway Explorer* are described in Table 1.

As illustrated in Fig. 6, events were mined for a 1-year time period, beginning with the patient's diagnosis of hyperlipidemia. In order to determine the patient outcomes associated with treatments, we analyzed the 90-days following the 1-year time period, and computed the average LDL cholesterol levels according to the patient's lab results. As recent research suggests that the target LDL level should be ~100 mg/dl, we classified patients with an average LDL of 100 or less to be labeled as a positive outcome, and patients with an average over 100 to be labeled as a negative outcome.

Initially, patterns were mined from 14,036 patients with hyperlipidemia. This population was balanced, so 50% of these patients had positive outcomes. The patients had a total of 70,379 diagnosis events and 97,189 medication events during their first year after diagnosis. However, few sequence patterns were found, even when setting the support to a low threshold such as 0.005. The patterns that were found are displayed in Fig. 7. Dr. Sorrentino speculated that by including all patients that simply had a hyperlipidemia diagnosis, the diversity of the types and order of clinical events these patients had would vary greatly, and that interesting patterns might be lost due to the extremely low support.

However, Dr. Sorrentino hypothesized that if he focused on specific sub-cohorts with pre-conditions, the analysis might lead to additional insights. He decided to investigate two pre-conditions: hypertension and diabetes.

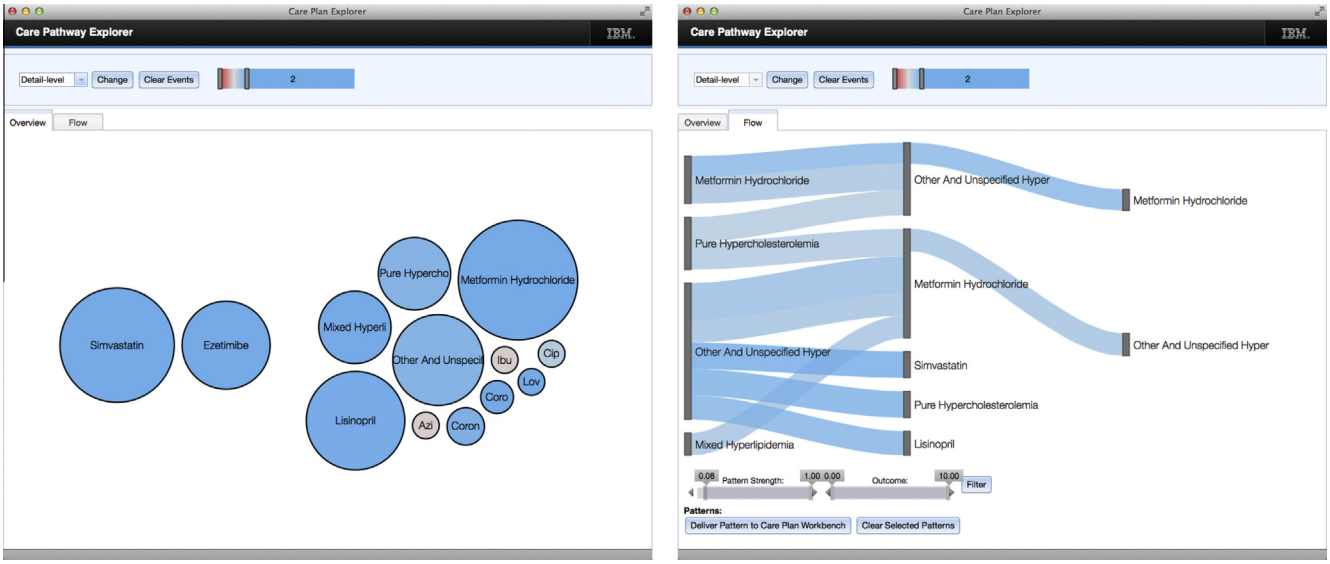

Fig. 4. Care Pathway Explorer features two complementary visualizations, an overview which resembles a bubble chart and displays events of the most frequent patterns mined (left), and a flow visualization to show the most frequent patterns (right).

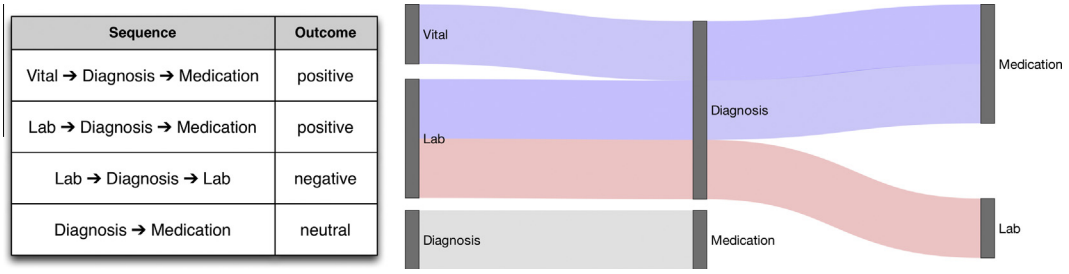

Fig. 5. An illustrative example of the flow visualization for a set of frequent patterns. Events in the frequent patterns are represented as nodes, and event nodes that belong to the same pattern are connected by edges. Patterns are colored according to their correlation with patient outcomes.

Table 1
List of the ICD-9 Groups used to mine diagnoses from patient records containing the primary hyperlipidemia diagnoses, and the complications that result from hyperlipidemia.

ICD-9 group	Description
272	Disorders of lipoid metabolism
362	Other retinal disorders
410	Acute myocardial infarction
411	Other acute and subacute forms of ischemic heart disease
412	Old myocardial infarction
413	Angina pectoris
414	Other forms of chronic ischemic heart disease
429	Ill-defined descriptions and complications of heart disease
433	Occlusion and stenosis of precerebral arteries
434	Occlusion of cerebral arteries
435	Transient cerebral ischemia
440	Atherosclerosis
996	Complications peculiar to certain specified procedures
V12	Personal history of certain other diseases

3.1. Hyperlipidemia with diabetes pre-condition

After filtering to patients with a diabetes pre-condition before the initial hyperlipidemia diagnosis, there were 1386 patients after balancing, featuring 11,058 hyperlipidemia-related diagnoses and 20,693 medication events.

Dr. Sorrentino was interested in analyzing patients with co-existing diabetes as they often have unhealthy lipid profiles with elevated LDL as a result of the changes to the metabolic pathways that channel the breakdown products of excess glucose into increased LDL production. Typically, prevention can be achieved through diet, exercise and HMG CoA Reductase Inhibitor usage to keep LDL less than 100 mg/dl. Thus, he was reassured to see that 'Dyslipidemics, HMG CoA Reductase Inhibitor' was a common event in most of the blue patterns at the top of Fig. 8.

After using Care Pathway Explorer to interactively filter to include only negative patterns, Dr. Sorrentino noticed that most of the negative patterns are related to drug-related side effects. Most of the drugs in question involved increases in LDL production

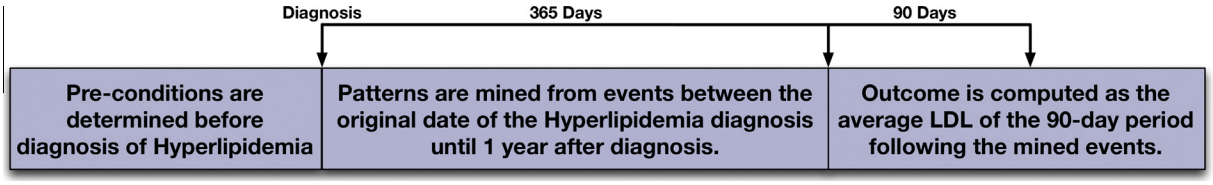

Fig. 6. In the long-term case study, events were mined for a 1-year time period, beginning with the patient's diagnosis of hyperlipidemia. In order to determine the patient outcomes associated with treatments, the 90-days following the 1-year time period were analyzed to compute the average LDL cholesterol levels according to the patient's lab results.

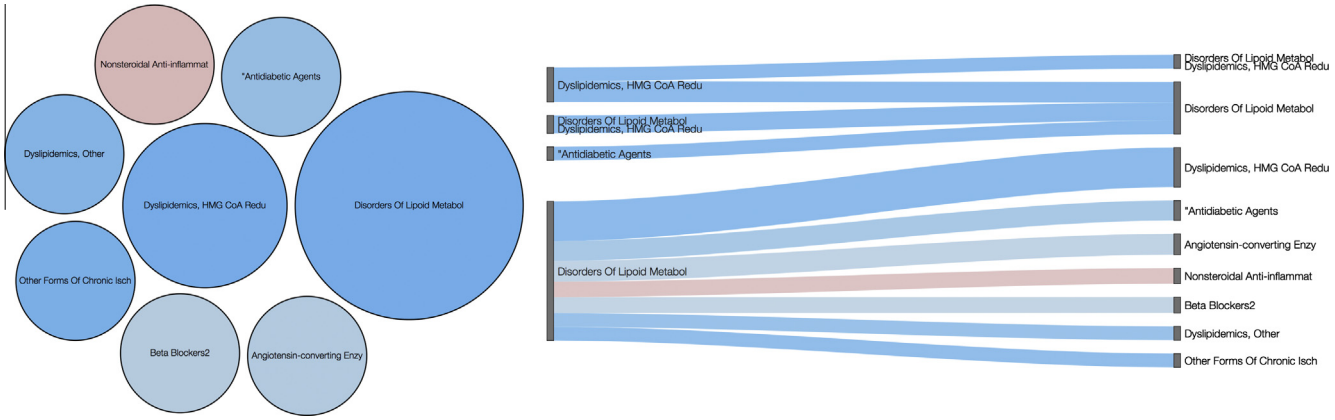

Fig. 7. Patterns mined using Care Pathway Explorer on a population of 14,036 patients with hyperlipidemia, who had a total of 70,379 diagnosis events and 97,189 medication events during their first year after diagnosis. This figure illustrates that few sequence patterns were found, even when setting the support to a low threshold such as 0.005.

or decreases in LDL removal; both lead to increases in LDL level. He noted that it is well documented in the medical literature that several antibiotic groups, including macrolides such as Azithromycin, and fluoroquinolones such as Ciprofloxacin, can increase LDL levels. Finally, glucocorticoids such as Cortisone and Prednisone also increase LDL levels, as do many oral contraceptives containing progesterone-related hormones.

3.2. Hyperlipidemia with hypertension pre-condition

After filtering to patients with a hypertension pre-condition before the initial hyperlipidemia diagnosis, there were 2800 patients after balancing, featuring 14,979 hyperlipidemia-related diagnoses and 24,898 medication events.

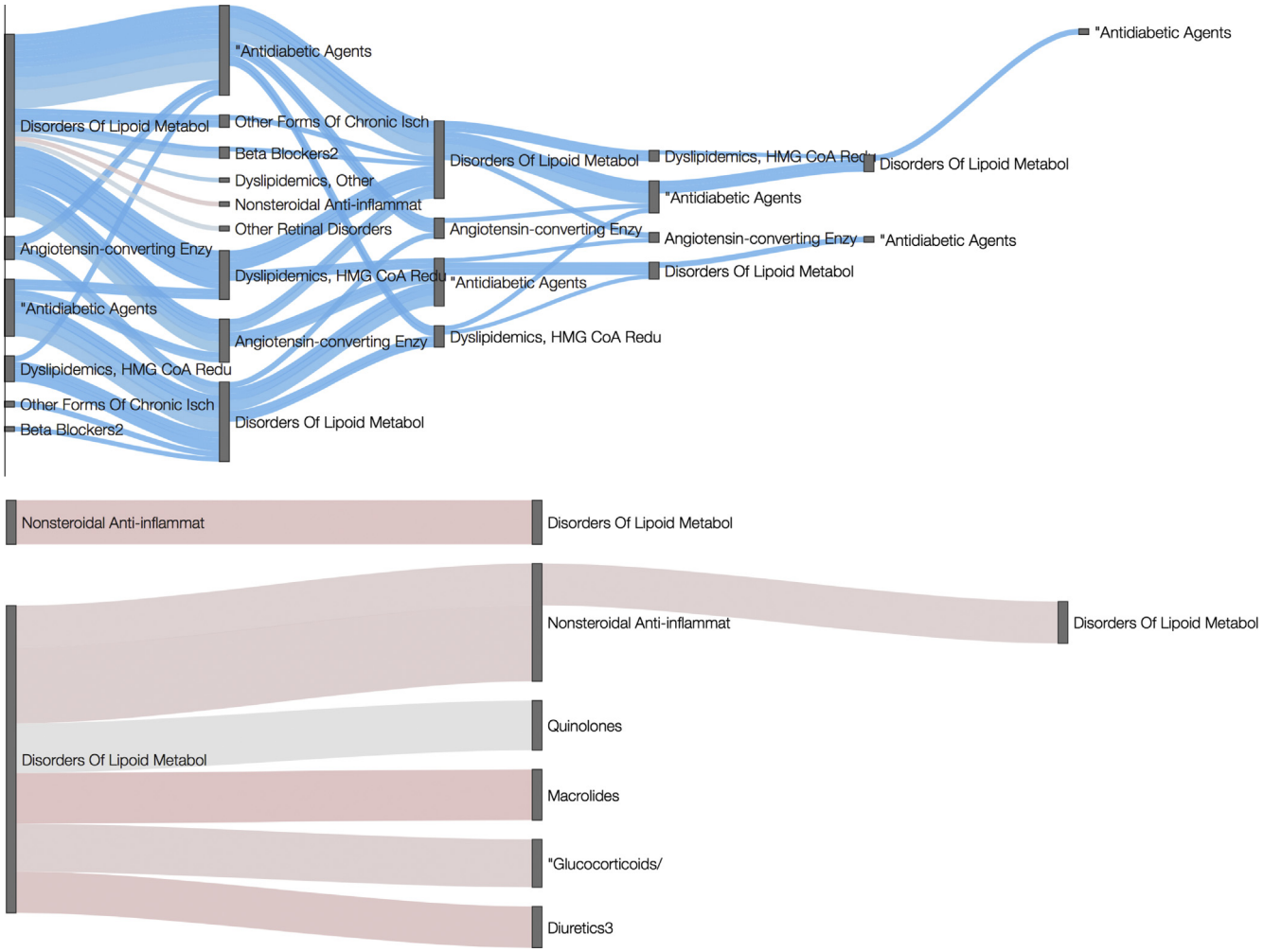

Fig. 8. Examples of insights reached interactively when mining patterns from patients with a diabetes pre-condition before the initial hyperlipidemia diagnosis (1,386 patients, featuring 11,058 hyperlipidemia-related diagnoses and 20,693 medication events).

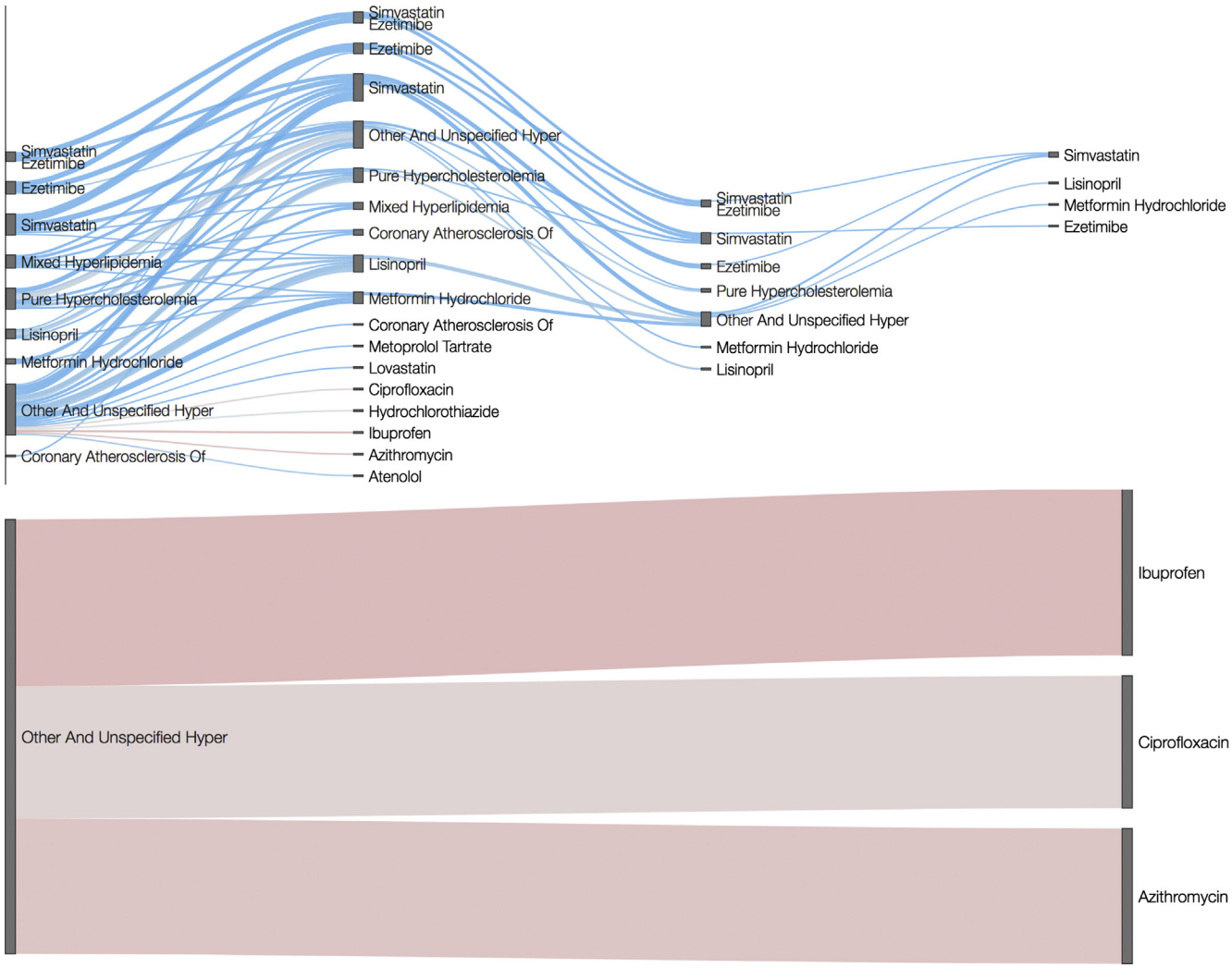

Fig. 9. Examples of insights reached interactively when mining patterns from patients with a hypertension pre-condition before the initial hyperlipidemia diagnosis (2,800 patients, featuring 14,979 hyperlipidemia-related diagnoses and 24,898 medication events).

The patterns at the finest level-of-detail are shown at the top of Fig. 9 using a minimum support threshold of 0.03. Here, these patterns show events at the ingredient level (e.g. Simvastatin, Ezetimibe) as well as detailed diagnoses (e.g. Pure Hypercholesterolemia, Coronary Atherosclerosis). Like the patterns common to patients with the Diabetes pre-condition, many patterns are associated with positive outcomes. Focusing on negative patterns, it is clear there are also negative side-effects of medications. However, while macrolides, such as azithromycin, were also present in the diabetes cohort, the hypertension cohort also features fluoroquinolones, such as Ciprofloxacin, which medical literature suggests can also increase LDL levels. There was no clear clinical link to the use of Ibuprofen among negative outcome patients, but it led to speculation that perhaps patients with higher LDL cholesterol levels also tend to use higher levels of pain medication.

4. Conclusions

As mentioned, some of the patterns were initially surprising to the clinical researcher. He suspects that many of the prescribers of certain medications are not aware that there may be side effects associated with raising LDL levels. In fact, only after examining medical literature was Dr. Sorrentino able to confirm the link between treatments featured in the negative patterns and higher LDL levels.

While this provides evidence that Care Pathway Explorer leads to insights, there remains future work to address certain limitations. For instance, there are still scalability issues with the mining algorithm. Even though our approach uses a computationally efficient bitmap-approach, large datasets with many concurrent events will slow down the algorithm. We are investigating the possibility to deploy the Frequent Pattern Analytics in a cloud-based distributed architecture, which is compatible with the hierarchical aspects of the algorithm. Addressing this scalability issue will also help determine the efficacy of our approach in settings where real-time decision support is needed.

Another issue is that the analytics currently require manual specification of certain parameters, such as the support threshold. While users can interactively specify these parameters, users must wait for the analytics to finish before finding out how many patterns the parameters result in. We plan to improve this by providing scented widgets [32] that inform the users how many patterns their choices may lead to before they commit to a specification.

While we have provided a long-term case study of the system, additional validation is required to fully understand the implications of mining care pathways and for determining how such techniques can be used for more than validating known findings. We

plan to deploy and evaluate the system with more datasets and case studies to better understand the role of these technologies in clinical decision support.

Another potential issue is that clinical data in practice may contain redundant or erroneous data that may have implications on the resulting analysis [38]. While addressing these issues is beyond the scope of our paper, Care Pathway Explorer has been tested with real-world EMR data with success. Furthermore, tools to interactively mine and visualize EMR data may expose data quality issues that might have otherwise left hidden in the database. Care Pathway Explorer also currently focuses on finding patterns among clinical events of a temporal nature. However, there is an opportunity to determine how these analytics can be expanded to handle other types of patient information (e.g. age, gender, ethnicity) are needed to see if stronger correlations could be found for specific patient profiles.

Despite these current limitations, our results suggest that *Care Pathway Explorer*, which automatically combines frequent sequence mining techniques with advanced visualizations, supports the integration of data-driven insights into care plan templates. As institutions evolve to handle the creation of more personalized care plan templates, these techniques will be valuable to ensure those new templates reflect best practices learned from past treatment actions and associated outcomes on similar patients.

Funding

This work was funded entirely by IBM. In addition, all authors are IBM employees and received a salary from the IBM Corporation.

Conflict of interest

The authors have no competing interests to declare.

Contributorship statement

AP designed and implemented the Care Pathway Explorer, conducted the long-term case study, analyzed the data, and drafted and revised the paper. He is guarantor. FW implemented the Frequent Sequence Mining analytics, analyzed the data, and drafted and revised the paper. JH analyzed the data and drafted and revised the paper.

Acknowledgment

We thank Dr. Robert Sorrentino for participating in our case study, and Ping Zhang for assistance with the preparation of data.

References

[1] T. Mitsa, Temporal Data Mining, first ed., Chapman & Hall/CRC, 2010.
[2] Z. Huang, X. Luemail, H. Duan, On mining clinical pathway patterns from medical behaviors, Artif. Intell. Med. 56 (1) (2012) 35–50.
[3] H. Scheuerlein, F. Rauchfuss, Y. Dittmar, R. Molle, T. Lehmann, N. Pienkos, et al., New methods for clinical pathways – business process modeling notation (BPMN) and tangible business process modeling (t.BPM), Langenbecks Arch. Surg. 397 (5) (2012) 755–761.
[4] W. Yao, A. Kumar, CONFlexFlow: integrating flexible clinical pathways into clinical decision support systems using context and rules, Decis. Support Syst. (2012) 499–515.
[5] G. Norén, J. Hopstadius, A. Bate, K. Star, I. Edwards, Temporal pattern discovery in longitudinal electronic patient records, Data Min. Knowl. Discov. 20 (3) (2010) 361–387.
[6] L. Chittaro, C. Combi, Visualizing queries on databases of temporal histories: new metaphors and their evaluation, Data Knowl. Eng. 44 (2) (2003) 239–264.
[7] J. Fails, A. Karlson, L. Shahamat, B. Shneiderman, A visual interface for multivariate temporal data: finding patterns of events across multiple histories, IEEE Symp. Vis. Anal. Sci. Technol. (2006) 167–174.
[8] C. Plaisant, S. Lam, B. Shneiderman, M.S. Smith, D. Roseman, G. Marchand, et al., Searching electronic health records for temporal patterns in patient histories: a case study with Microsoft Amalga, AMIA Annu. Symp. Proc. (2008) 601–605.
[9] F. Mörchen, A. Ultsch, Efficient mining of understandable patterns from multivariate interval time series, Data Min. Knowl. Discov. 15 (2) (2007) 181–215.
[10] Y. Shahar, A framework for knowledge-based temporal abstraction, Artif. Intell. 90 (1997) 79–133.
[11] Y. Shahar, D. Goren-Bar, D. Boaz, G. Tahan, Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions, Artif. Intell. Med. 38 (2006) 115–135.
[12] Y. Shahar, D. Goren-Bar, D. Boaz, G. Tahan, Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions, Artif. Intell. Med. 38 (2) (2006) 115–135.
[13] D. Klimov, Y. Shahar, M. Taieb-Maimon, Intelligent visualization and exploration of time-oriented data of multiple patients, Artif. Intell. Med. 49 (1) (2010) 11–31.
[14] A. Bertone, T. Lammarsch, T. Turic, W. Aigner, S. Miksch, J. Gaertner, MuTIny: a multi-time interval pattern discovery approach to preserve the temporal information in between, in: Proceedings of the IADIS European Conference on Data Mining, 2010, pp. 101–106.
[15] T. Lammarsch, W. Aigner, A. Bertone, S. Miksch, A. Rind, Mind the time: unleashing temporal aspects in pattern discovery, Comput. Graphics 38 (2014) 38–50.
[16] T. Gschwandtner, W. Aigner, K. Kaiser, S. Miksch, A. Seyfang, CareCruiser: exploring and visualizing plans, events, and effects interactively, in: Proceedings of the IEEE Pacific Visualization Symposium, 2011, pp. 43–50.
[17] K. Wongsuphasawat, J.A. Guerra Gomez, C. Plaisant, T.D. Wang, M. Taieb-Maimon, B. Shneiderman, Lifeflow: visualizing an overview of event sequences, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 1747–1756.
[18] M. Monroe, R. Lan, C. Plaisant, B. Shneiderman, Temporal event sequence simplification, IEEE Trans. Visualization Comput. Graphics 19 (12) (2013) 2227–2236.
[19] K. Wongsuphasawat, D. Gotz, Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization, IEEE Trans. Visualization Comput. Graphics 18 (12) (2012) 2659–2668.
[20] Adam Perer, Fei Wang, Frequence. Interactive mining and visualization of temporal frequent event sequences, in: ACM Conference on Intelligent User Interfaces, 2014, pp. 153–162.
[21] D. Gotz, H. Stavropoulos, J. Sun, F. Wang, ICDA: a platform for intelligent care delivery analytics, AMIA Annu. Symp. Proc. (2012) 264–273.
[22] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, VLDB (1994) 487–499.
[23] J. Ayres, J. Flannick, J. Gehrke, T. Yiu, Sequential PAttern mining using a bitmap representation, in: Proc Eighth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2002, pp. 429–35.
[24] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (6) (2004) 066133.
[25] W. Wang, H. Wang, G. Dai, H. Wang, Visualization of large hierarchical data by circle packing, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2006, pp. 517–520.
[26] P. Riehmann, M. Hanfler, B. Froehlich, Interactive Sankey diagrams, IEEE Inf. Visualization Symp. (2005) 233–240.
[27] Yiqin Yu, Haifeng Liu, Jing Li, Xiang Li, Jing Mei, Guotong Xie, Adam Perer, Fei Wang, Jianying Hu, Care pathway workbench: evidence harmonization from guideline and data, in: European Medical Informatics Conference, 2014.
[28] E. Bertini, H. Lam, A. Perer, Summaries: a special issue on evaluation for information visualization, Inf. Visualization 10 (3) (2011).
[29] C. Plaisant, The challenge of information visualization evaluation, in: Proceedings of the Working Conference on Advanced Visual Interfaces, 2004, pp. 109–116.
[30] B. Shneiderman, C. Plaisant, Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies, in: Proceedings of the BELIV Workshop, 2006, pp. 1–7.
[31] A. Perer, B. Shneiderman, Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 265–274.
[32] W. Willett, J. Heer, M. Agrawala, Scented widgets: improving navigation cues with embedded visualizations, IEEE Trans. Visual Comput. Graphics 13 (6) (2007) 1129–1136.
[33] L. Sacchi, A. Dagliati, R. Bellazzi, Analyzing complex patients' temporal histories: new frontiers in temporal data mining, Data Min. Clin. Med. 1246 (2015) 89–105.
[34] Denis Klimov, Alexander Shknevsky, Yuval Shahar, Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory, J. Am. Med. Inf. Assoc. (2014), http://dx.doi.org/10.1136/amiajnl-2014-002927.
[35] R. Moskovitch, Y. Shahar, Fast time intervals mining using the transitivity of temporal relations, Knowl. Inf. Syst. 42 (1) (2015) 21–48.

[36] Iyad Batal, Gregory F. Cooper, Dmitriy Fradkin, James Harrison Jr., Fabian Moerchen, Milos Hauskrecht, An efficient pattern mining approach for event detection in multivariate temporal data, Knowl. Inf. Syst. (2015) 1–36.
[37] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Ben Shneiderman, Cohort comparison of event sequences with balanced integration of visual analytics and statistics, in: Conference on Intelligent User Interfaces, 2015.
[38] S. Bowman, Impact of electronic health record systems on information integrity: quality and safety implications, Perspect. Health Inf. Manage. 10 (Fall) (2013) 1c.
[39] Dmitriy Fradkin, Fabian Mörchen, Mining sequential patterns for classification, Knowl. Inf. Syst. (2015) 1–19.
[40] Robert Moskovitch, Yuval Shahar, Classification-driven temporal discretization of multivariate time series, Data Min. Knowl. Disc. (2014) 1–43.
[41] Robert Moskovitch, Yuval Shahar, Classification of multivariate time series via temporal abstraction and time intervals mining, Knowl. Inf. Syst. (2014) 1–40.
[42] Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, Milos Hauskrecht, A temporal pattern mining approach for classifying electronic health record data, ACM Tran. Intell. Syst. Technol. (TIST) 4 (4) (2013) 63.
[43] Dhaval Patel, Wynne Hsu, Mong Li Lee, Mining relationships among interval-based events for classification, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp. 393–404.