Chapter 31
Data Driven Analytics for Personalized Healthcare

Jianying Hu, Adam Perer, and Fei Wang

Abstract The concept of Learning Health Systems (LHS) is gaining momentum as more and more electronic healthcare data becomes increasingly accessible. The core idea is to enable learning from the collective experience of a care delivery network as recorded in the observational data, to iteratively improve care quality as care is being provided in a real world setting. In line with this vision, much recent research effort has been devoted to exploring machine learning, data mining and data visualization methodologies that can be used to derive real world evidence from diverse sources of healthcare data to provide personalized decision support for care delivery and care management. In this chapter, we will give an overview of a wide range of analytics and visualization components we have developed, examples of clinical insights reached from these components, and some new directions we are taking.

Keywords Data driven healthcare analytics • Learning health system • Practice based evidence • Real world evidence • Clinical decision support • Machine learning • Data mining • Data visualization

31.1 Introduction

In recent years we have witnessed a dramatic increase of electronic health data, including extensive Electronic Medical Records (EMR) recording patient conditions, diagnostic tests, labs, imaging exams, genomics, proteomics, treatments, outcomes, claims, financial records, clinical guidelines and best practices etc.

J. Hu, PhD (✉) • A. Perer, PhD
Healthcare Analytics Research Group, IBM T.J. Watson Research Center,
Yorktown Heights, NY, USA
e-mail: jyhu@us.ibm.com; adam.perer@us.ibm.com

F. Wang, PhD
Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

C.A. Weaver et al. (eds.), *Healthcare Information Management Systems: Cases, Strategies, and Solutions*, Health Informatics,
DOI 10.1007/978-3-319-20765-0_31

Healthcare professionals are now increasingly asking the question: what can we do with this wealth of data? How can we perform meaningful analytics on such data to derive insights to improve quality of care and reduce cost?

Healthcare Analytics needs to cover the whole spectrum including both Knowledge Driven Analytics and Data Driven Analytics. Knowledge driven approaches operate on knowledge repositories that include scientific literature, published clinical trial results, medical journals, textbooks, as well as clinical practice guidelines. Traditionally the gold standard of evidence in healthcare has been produced through the randomized controlled trial process. Results of such trials get published and then healthcare professionals consult those publications to bring to the point of care nuggets of evidence that apply to the scenario at hand. This process of knowledge diffusion can take as long as 17 years [2, 10].

Innovations such as Watson Discovery Advisor [32] can dramatically reduce that time frame to close the knowledge diffusion gap. In tools such as WDA, the deep NLP and Q&A capabilities such those developed in the original Watson Jeopardy machine are leveraged to teach the computer to learn medical domain knowledge from unstructured data captured in the knowledge repository, and then make intelligent inference from such knowledge to bring the most relevant pieces of information to the finger tips of the practitioners.

Complementary to this knowledge dissemination processing, data driven healthcare analytics is about making the computer learn from observational data collected in the process of delivering care. This is important because published guidelines typically target a single disease and the average patient, so by themselves don't provide sufficient insight into how to best manage a real world patient with multiple comorbidities and complex conditions. By tapping into the vast real world observational data collected at the individual patient level, we can leverage the collective experience of a healthcare delivery system, to extract insights that can be used to fill in that personalization gap, and in that process continuously enhance and refine our knowledge on best practices. Such insights are referred to as Practice Based Evidence, or Real World Evidence, and are at the center of the vision of "Learning Health Systems" advocated by the Institute of Medicine [11].

The focus of this chapter is on data driven healthcare analytics. In line with the vision of LHS, the healthcare analytics research group at IBM has been working on applying advanced machine learning, data mining and data visualization techniques in the context of real world healthcare data and use cases to build up a data driven healthcare analytics framework. An earlier version of this analytics framework, called Intelligent Care Delivery Analytics (ICDA), was reported in AMIA 2012 [6]. We have continued to expand this framework by adding more machine learning and visualization components since then, and expect to continue to do so in the future.

Figure 31.1 gives an overview of the current snapshot of ICDA. At the center of this framework is Patient Similarity Analytics. The objective of this component is to develop methodologies that can be used to identify patients who are similar to a patient of interest in a clinically meaningful way, so that insights derived from longitudinal records of the similar patients can be used to help determine personalized

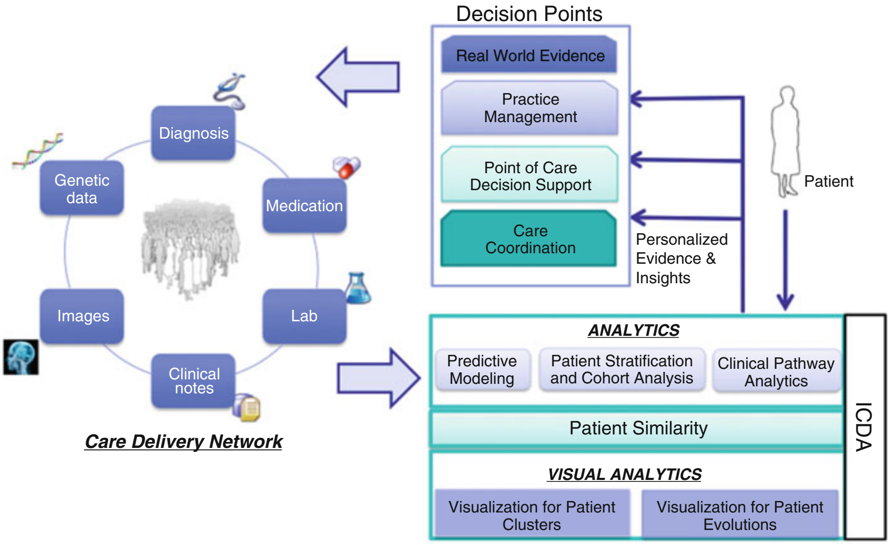


Fig. 31.1 Intelligent Care Delivery Analytics (ICDA) – the data driven healthcare analytics platform at IBM research

prognosis and treatment plans for this specific patient. Building on and around this central component, we have developed a suite of analytics and visualization components to address challenges and use cases encountered in different aspects of the care process, and deliver insights in an interactive, consumable manner.

Throughout the rest of the chapter, we will describe some of the key elements, and provide concrete examples of novel algorithms that have been developed in this framework. Due to the limited space, we focus on describing the functionalities and high-level approaches of these analytics and visualization components. In-depth technical details and discussions can be found in the numerous publications cited throughout the chapter.

31.2 Patient Similarity Analytics

Existing EMR systems typically store data in a manner that makes it difficult for clinicians to extract what is necessary to make clinical decisions at the point-of-care. Most of EMR systems are primarily used to record clinical events for bookkeeping and claim purposes as opposed to be used as a decision support tool for better diagnosis and treatment. Constructing a patient network with nodes representing patients and edges connecting clinically similar patients could be very helpful to such a clinical decision support system, as the physician can look at the treatments and disease condition evolutions of the similar patients to come up with a better care plan for the current patient.

Besides decision support systems, there are also other areas in medical informatics where such patient network could be very helpful. For example, Comparative Effectiveness Research [1], which is the direct comparison of existing health care interventions to determine which work best for which patients and which pose the greatest benefits and harms [w1]. In this case, if we can first stratify the patients into different cohorts according to their clinical similarity, then CER can be performed on the patients within the same cohorts [13]. Under a similar setting, patient risk stratification aims to stratify the patients according to their disease condition risks. This is a crucial step for effective management of patients because for patients with different risks, we may have different treatment plans. Furthermore, if we can construct an undirected patient network using such patient similarity, we can expect to discover clinically meaningful insights such as disease evolution patterns and care or treatment patterns.

31.2.1 Patient Similarity Metric Learning

While traditional patient cohort generation tools such as i2b2 [16] address some aspects of patient similarity, they are limited in that cohorts have to be identified through database queries using a few pre-selected attributes. To fully realize the power of patient similarity analytics, a big data approach is needed where all known attributes about patients are taken into consideration, in order to account for all potential confounding factors. This poses two challenges. First, since the number of attributes can be very large (e.g., in the order of tens of thousands), how to define distance, or similarity metric, in this high dimensional space is a challenging mathematical problem. Second, the notion of patient similarity is context dependent. For example, the factors that are important for identifying similar patients in the context of determining best treatment for hyperlipidemia may be completely different from the ones for evaluating different chemotherapies for a cancer patient. To address these challenges, machine learning approaches called metric learning are needed to derive from data the most appropriate similarity metric, i.e., most important attributes along with the weighting factors for a specific clinical context.

Patient EMRs contain a large amount of features coming from heterogeneous sources, such as demographic information, diagnosis, medication, lab tests and so on and so forth. To facilitate the process of similarity learning, researchers have proposed constructing a profile for each patient, which is a feature vector with the dimensionality equal to the number of different features. Before constructing such a vector, a time period of interest is defined, within which the features are aggregated to obtain the entries in the patient profile (e.g., the average value of a specific lab test, or the count of a specific diagnosis code). In this way, after profiling, each patient is represented as a feature vector [27, 28].

Local Supervised Metric Learning (LSML) is a supervised metric learning approach that has been proved to be useful in patient similarity evaluation [3, 23, 24].

This algorithm was initially proposed in [30] for face recognition. The basic idea of LSML is to maximize the local separability of the data vectors from different classes.

We applied the LSML in the context of monitoring patients in the Intensive Care Unit (ICU) [3]. ICUs are data rich environments where patients are continuously being monitored for several aspects of their health. Alerts that can indicate the likely onset of an imminent adverse condition based on the behavior of patients' temporal data provide important support mechanism for physicians in this environment. Accompanying those alerts with insight regarding the likely behavior of patient KPIs can further qualify and clarify them. In this setting, our goal is to retrieve patients who display similar evolution patterns in their ICU data to the patient being monitored and use the future trend of the cohort of similar patients to predict if the patient being monitored is going to experience a medical event within a specific time horizon. The insight provided to the clinician through the projections of the patient's physiological data into the future could further clarify and qualify the generated alerts. The proposed approach and system were tested using the MIMIC II database, which consists of physiological waveforms, and accompanying clinical data obtained for ICU patients.

31.2.2 Inference Over Multiple Similarity Networks for Personalized Medicine

Using patient similarity analytics methods such as the described above, one can construct patient similarity networks where each node represents a patient and the edge between a pair of patients represent the degree of similarity between the two patients represented by their key clinical indicators. In a recent work, we augmented this patient similarity network with a drug similarity network, and developed a machine learning approach to make inferences over this heterogeneous network to derive Real World Evidence for personalized drug response prediction [36]. To compose the drug similarity network, we used chemical structure extracted from PubChem, and drug target protein information extracted from DrugBank [34]. Links between patients and drugs were then constructed to represent the prior associations between patients and drugs, which were measured by the Tonimoto Coefficient between ICD9 diagnosis codes of patients and ICD9-format drug indications from MEDI database [33]. Finally, a machine learning technique called label propagation [31] was applied to infer, for any given drug, the likely effectiveness of this drug on any of the patients who have not yet received this drug. Intuitively, this allows us to infer the likely response of a patient to a particular drug based on observations of how similar patients have responded to similar drugs in the past. Experimental evaluation results on a real-world EMR dataset of 110,157 hyperlipidemia patients demonstrate the effectiveness of the proposed method and suggest that the combination of appropriate patient similarity and drug similarity analytics can help identify which drug is likely to be effective for a given patient.

31.2.3 CareFlow: Data-Driven Visual Exploration of Similar Patients

Patient similarity analytics can be combined with advanced visualization techniques to provide physicians with the most relevant information in a consumable manner. One example of such a tool is CareFlow, for the exploration of care pathways from similar patients. When a patient is diagnosed with a disease, their doctor will often devise a care pathway, a sequence of medical treatments to help manage their disease or condition. When doctors devise care pathways, they often must rely on their education, experience, and intuition [25]. The goal of CareFlow [17] is to leverage the rich longitudinal data found in Electronic Medical Records (EMRs) to empower clinicians with a new data-driven resource for the design of personalized care plans. Using the relevant clinical data of a specific patient, CareFlow mines the EMRs to find clinically similar patients using our patient similarity analytics. CareFlow then visualizes all of the different care pathways that these similar patients have undergone, while providing context on which care pathways were successful and which were not. The resulting visualization supports the identification of the most desirable and most problematic care plans.

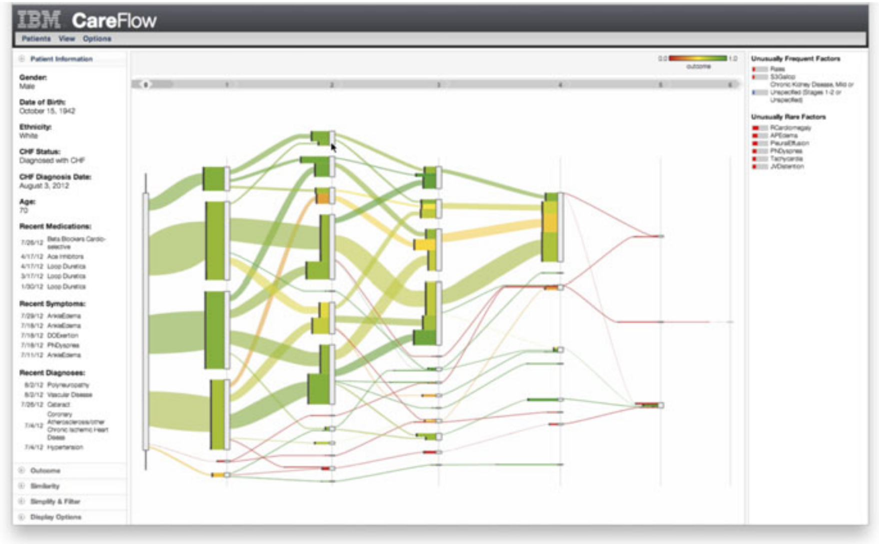
31.2.4 Mining Care Pathways from Data

In order to model the care pathways for the similar patient population, CareFlow mines the EMRs for relevant patient events. For each similar patient, CareFlow will extract records of performed treatments and their associated dates by querying the EMR database for relevant medical events. The result of this query is a complex dataset describing the details of various treatments given to the entire similar patient population.

Of course, each similar patient underwent treatments at different points in time. In order to unify them, CareFlow aligns all treatments by the time at which each patient was first diagnosed with the disease of interest. CareFlow defines the care pathway as the sequence of treatments after diagnosis. In addition to deriving care pathways, outcomes are also derived from the EMRs for each of these similar patients. By associating each care pathway with an outcome, it is possible to infer which care pathways lead to statistically better outcomes. CareFlow makes this outcome information visually prominent to inform medical decisions.

31.2.5 Visualizing Care Pathways

While a doctor may be able to make sense of a care plan for a single patient (e.g. [20]), doing so for a similar patient population is much more challenging. Care pathways may have a large number of different types of treatments, and the sequence



Gender:
Male

Date of Birth:
October 15, 1942

Ethnicity:
White

CHF Status:
Diagnosed with CHF

CHF Diagnosis Date:
August 3, 2012

Age:
70

Recent Medications:

7/26/12	Beta Blockers Cardio-selective
4/17/12	Ace Inhibitors
4/17/12	Loop Diuretics
3/17/12	Loop Diuretics
1/30/12	Loop Diuretics

Recent Symptoms:

7/29/12	AnkleEdema
7/18/12	AnkleEdema
7/18/12	DOExertion
7/18/12	PNDyspnea
7/11/12	AnkleEdema

Recent Diagnoses:

8/2/12	Polyneuropathy
8/2/12	Vascular Disease
7/26/12	Cataract
7/4/12	Coronary Atherosclerosis/other Chronic Ischemic Heart Disease
7/4/12	Hypertension

Unusually Frequent Factors
- Rales
- S3Gallop
- Chronic Kidney Disease, Mild or Unspecified (Stages 1-2 or Unspecified)

Unusually Rare Factors
- RCardiomegaly
- APEdema
- PleuralEffusion
- PNDyspnea
- Tachycardia
- JVDistention

Fig. 31.2 CareFlow's visual interface. The *left panel* displays a summary of the patient's relevant medical history. The *center panel* displays a visualization of the care plans of the 300 most similar patients. The *right panel* displays the factors associated with a selected subset of patients

of these treatments often varies as well. CareFlow provides a visualization of the temporal sequence of treatments. As shown in Fig. 31.2, treatments are represented as nodes and positioned along the horizontal axis, which represents the sequence of treatments over time. The diagnosis of a disease occurs on the far left of the visualization, and treatments in the care plan extend to the right. The height of each node is proportional to the number of patients that took a given treatment. Link edges are also present to connect nodes from their previous and future nodes in the care pathway. The visual elements are colored according to the average outcome of all patients represented by the node or edge. Elements that are colored green represented parts of the care plan where patients remained healthy, whereas elements that are colored red indicate care plans of patients who ended up in poor health.

31.2.6 Use Case: Congestive Heart Failure

This use case involves a doctor who has recently diagnosed a patient with congestive heart failure and wishes to use CareFlow to examine the historical outcomes of possible care pathways. CareFlow connects to a longitudinal EMR database of over 50,000 patients with heart conditions spanning over 8 years.

On the left-hand side of Fig. 31.2, a summary of the patient's relevant medical history is shown, including recent medications, symptoms, and diagnoses. In the center panel of Fig. 31.2, a visualization of the care plans of the 300 most similar patients is shown. The left-most node represents these similar patients at their point of diagnosis with heart failure. As the visualization extends to the right, the various treatment sequences of similar patients are shown. The care pathways are colored according to a continuous color scale,; pathways that are colored red implies most patients within that node ended up being hospitalized, whereas green pathways means most patients managed to stay out the hospital.

In addition to gaining an overview of all care pathways, a doctor can also focus on the most successful treatment plan. By selecting the appropriate button, the care plan that leads to the best outcomes for patients is highlighted.

CareFlow provides doctors with the ability to get more information about the patients who undertook a particular care plan. By selecting a Treatment node, doctors can view a precise count of the number of patients the node represents, as well as the average outcome for these patients. In addition, the right panel of the interface displays summary information about a set of patients by displaying factors common to this cohort, as well as factors rare in this group.

31.3 Predictive Modeling

Healthcare analytics research increasingly involves the construction of predictive models for disease targets across varying patient cohorts using observational data such as EMR. A common workflow for predictive models is a five-step process: (1) cohort construction, (2) feature engineering, (3) cross-validation, (4) feature selection, and (5) classification/model selection. We have developed novel machine learning and visualization methods to help address the challenges faced in each of these steps.

31.3.1 Feature Engineering

Feature Engineering, which is about inferring phenotypic patterns from population-scale clinical data, is a core computational task in the development of personalized medicine. One important source of data on which to conduct this type of research are patient EMRs. However, the patient longitudinal EMRs are typically sparse and noisy, which creates significant challenges if we use them directly to represent patient phenotypes. We developed a data driven phenotyping framework called Pacifier (PAtient reCord densIFIER) [37], where we interpret the longitudinal EMR data of each patient as a sparse matrix with a feature dimension and a time dimension, and derive more robust patient phenotypes by exploring the latent structure of those matrices. Specifically, we assume that each derived phenotype is composed of a subset of the medical features contained in original patient EMR, whose value evolves smoothly over time. We propose two formulations to achieve such goal. One is Individual Basis Approach (IBA), which assumes the phenotypes are different for every patient. The other is Shared Basis Approach (SBA), which assumes the patient population shares a common set of phenotypes. We developed an efficient optimization algorithm that is capable of resolving both problems efficiently. Pacifier was validated on two real world EMR cohorts for the tasks of early prediction of Congestive Heart Failure (CHF) and End Stage Renal Disease (ESRD). Our results showed that the predictive performance in both tasks can be improved significantly by the proposed algorithms (average AUC score improved from 0.689 to 0.816 on CHF, and from 0.756 to 0.838 on ESRD respectively).

31.3.2 Large Scale Feature Selection Algorithms

Another key challenge in developing risk prediction models from observational healthcare data is how to effectively identify, form the larger number (typically thousands to tens of thousands) of features the salient risk factors, i.e., the subset of features that are most predictive. Knowledge driven and data driven strategies reflect two ends of the spectrum of risk factor identification or feature selection. More specifically, a knowledge driven approach is based on evidence of varying quality, guidelines, and experts' opinions, while a data driven approach is solely based on the observational data. We developed a hybrid strategy that starts with prior knowledge, then extends to a more comprehensive model by selectively including an additional set of features that both optimize prediction and complement knowledge based features. In particular, we extended a sparse feature selection method called Scalable Orthogonal Regression (SOR) [12] to expand a set of knowledge driven risk factors with additional risk factors from data [22]. The method was designed specifically to select less redundant features without sacrificing the quality, for which redundancy is measured by an orthogonality measure added as a penalty term in the objective function. The approach was validated using a large dataset containing 4,644 heart failure cases and 45,981 controls. The proposed method was shown to identify complementary risk factors that are not in the existing known factors and can better predict the onset of HF. In other words, the combined risk factors between knowledge and data significantly outperform knowledge-based risk factors alone. Furthermore, those additional risk factors were confirmed to be clinically meaningful by a cardiologist [22].

31.3.3 Scalable Model Exploration

To develop an appropriate predictive model for healthcare applications, it is often necessary to compare and refine a larger number of models derived from a diversity of cohorts, patient-specific features, feature selection algorithms, and classifiers/regression methods. An efficient and scalable computing platform is required to facilitate such large scale models exploration. To support this goal, we developed a

PARAllel predictive MOdeling (PARAMO) platform [15] which (1) constructs a dependency graph of tasks from specifications of predictive modeling pipelines, (2) schedules the tasks in a topological ordering of the graph, and (3) executes those tasks in parallel. We implemented this platform using Map-Reduce to enable independent tasks to run in parallel in a cluster computing environment. Different task scheduling preferences are also supported.

We assessed the performance of PARAMO on various workloads using three datasets derived from the EMR systems in place at Geisinger Health System and Vanderbilt University Medical Center and an anonymous longitudinal claims database. We demonstrate significant gains in computational efficiency against a standard approach. In particular, PARAMO can build 800 different models on a 300,000 patient data set in 3 hours in parallel compared to 9 days if running sequentially.

This work demonstrates that an efficient parallel predictive modeling platform can be developed for EMR data. Such a platform can facilitate large-scale modeling endeavors and speed-up the research workflow and reuse of health information.

31.3.4 Visual Analytics for Predictive Modeling

When data is high-dimensional, feature selection algorithms are often used to remove non-informative features from models. Here the analyst is confronted with the decision of which feature selection algorithm to utilize, and even if the analyst decides to try out multiple types, the algorithmic output is often not amenable to user interpretation. This limits the ability for users to utilize their domain expertise during the modeling process. To improve on this limitation, INFUSE (INteractive FeatUre SElection) [9], was designed to help analysts understand how predictive features are being ranked across feature selection algorithms, cross-validation folds, and classifiers.

31.3.5 Use Case: Diabetes Prediction

In order to demonstrate the promise of visualizing predictive models, we describe an example scenario with clinical researchers interested in using predictive modeling on a longitudinal database of electronic medical records. Their database features over 300,000 patients from a major healthcare provider in the United States. The team is interested in building a predictive model to predict if a patient is at risk of developing diabetes, a chronic disease of high blood sugar levels that causes serious health complications.

From this database, the team constructs a cohort (Step 1) of patients. Fifty percent of these patients are considered incident cases with a diagnosis of diabetes. Each case was paired with a control patient based on age, gender, and primary care physician resulting in control patients without diabetes. From the medical records of these patients, they extract four meaningful types of features (Step 2): diagnoses, lab tests, medications, and procedures. Next, in order to reduce the bias of the predictive models, the team uses ten cross-validation folds (i.e. random samples) (Step 3) to divide the population randomly into ten groups. After cohorts, features, and folds are defined, the clinical researchers are ready to use feature selection. The team has four feature selection algorithms implemented and available to them (Step 4): these include Information Gain, Fisher Score, Odds Ratio and Relative Risk. Finally, the team evaluates each selected feature set as a model using four classifiers (Step 5): Logistic Regression, Decision Trees, Naive Bayes, and K-Nearest Neighbors.

Typically, this team executes a pipeline of multiple feature selection algorithms and chooses the model that ends up with the best scores from the classifier. However, while this approach allows the team to find the model with the highest accuracy score, they do not have direct access to view the features that make up the model. This is the goal of INFUSE: to make those features automatically selected more visible.

31.3.6 Visualizing Features

As described, the features are ranked by multiple feature selection algorithms and across multiple cross-validation folds. INFUSE's visual design embeds all of this information in a circular glyph that shows all the rankings obtained from each algorithm/fold pair. As shown in Fig. 31.2 INFUSE (a), the glyph is divided into equally-sized circular segments; where each segment represents one of the ranking algorithms. For instance, in Fig. 31.2 INFUSE (b), the feature was ranked by four feature selection algorithms, so the circular glyph is divided into four sections. These sections are then divided further into a fold slice for each cross-validation fold. For instance, in Fig. 31.2 INFUSE (c), each feature selection algorithm was executed on ten cross-validation folds, therefore there are tenfold slices. Within each fold slice, there is an inward-growing bar (that is, starting from the perimeter and growing towards the center) that represents the rank of the feature in a particular fold. For example, in Fig. 31.2 INFUSE (c), the feature is higher ranked in Fold 3 than in Fold 4 as the bar in Fold 3 stretches closer towards the center than in Fold 4. Features that are unranked, because their scores are too low to meet the minimum threshold requirement of the algorithm, are represented as empty slices with no bars.

The feature glyphs are displayed inside a zoomable visualization that allows users to find the features of interest. For instance they can group all features by type (e.g. medication, diagnosis, lab type, as seen in Fig. 31.3 INFUSE) or display the features on a scatterplot (e.g. average of features vs how many times a feature was picked by an algorithm, as seen in Fig. 31.4 INFUSE).

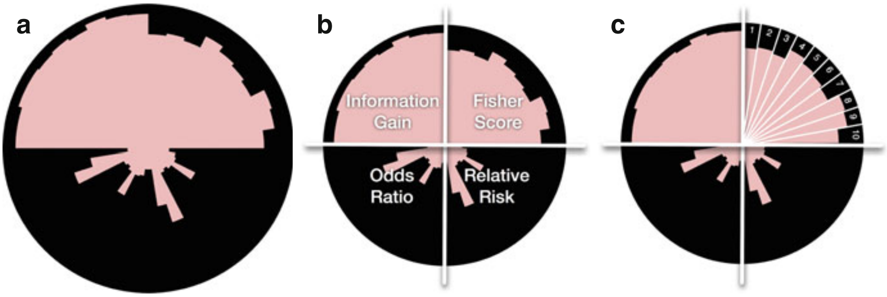


Fig. 31.3 (**a**) An illustration of how features are visually represented as circular glyphs. (**b**) Multiple models for each feature are represented as *model sections*. In this example, the feature is divided into four sections, as it was ranked by four feature selection algorithms (Information Gain, Fisher-Score, Relative Risk, and Odds Ratio.). (**c**) Each section is further divided into *fold slices* representing each of the cross-validation folds. Each fold slices features an inward-filling bar that represents the rank of this feature in that fold. A longer bar implies the feature has a better rank. If no bar appears, the feature was unranked in the fold, and thus did not meet the importance threshold

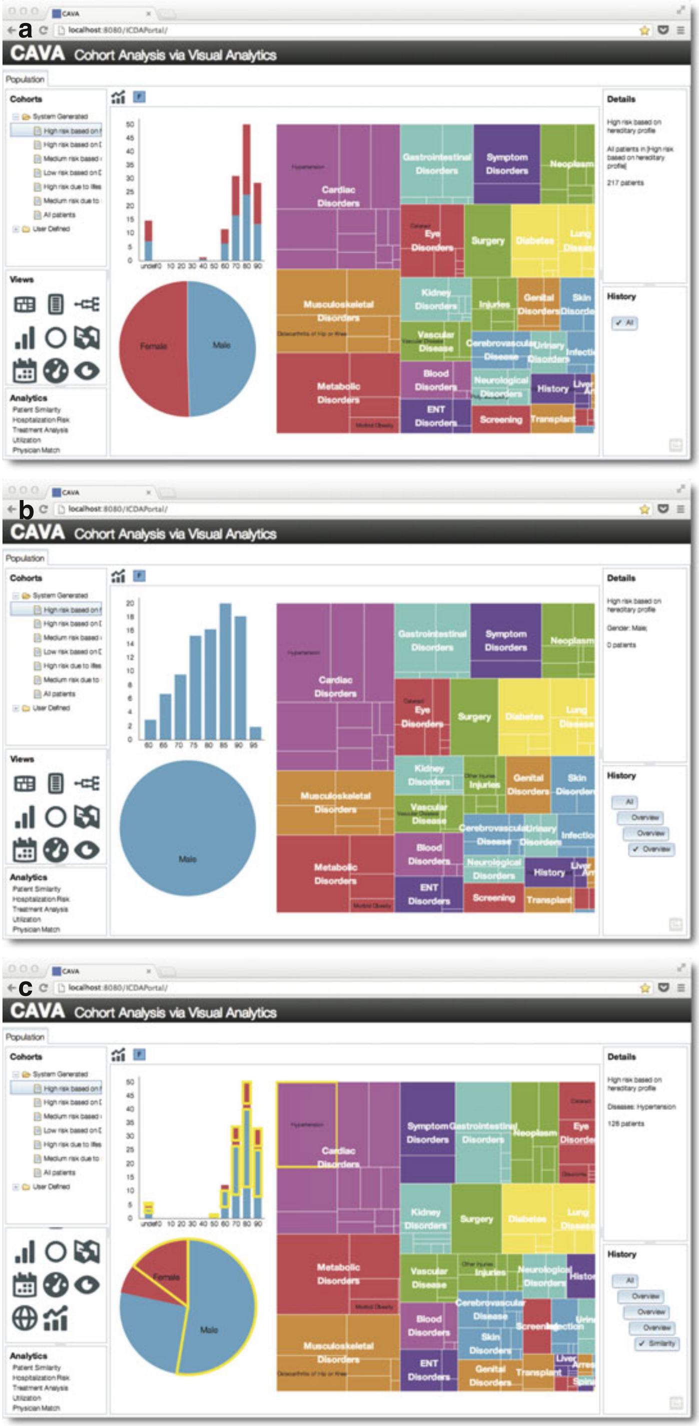
31.3.7 Finding Clinically Relevant Features

The following is an example of the types of insights that can be reached with INFUSE. When examining the scatterplot view (the top of Fig. 31.4 INFUSE), all of the medications that were ranked by all feature selection algorithms and folds and found that they were antihyperglycemic medications, which are common treatments to lower the blood sugar of diabetes patients, and made clinical sense to be ranked high.

However, looking towards the center of the scatterplot, where the features are only ranked by half of the algorithms and folds, it is noticeable that a cluster of medications that had half-circle patterns like those described above. This region is highlighted in the red box of Fig. 31.4 INFUSE. By mouse-hovering these features to read their names, it shows that those ranked high by the upper-half of the circle (Information Gain and Fisher Score) were as clinically relevant and similar as those ranked by the bottom-half algorithms (Relative Risk and Odds Ratio). This provided feedback that in predictive modeling it is not safe to assume that one single feature selection algorithm is able to detect all possible interesting features and also

→

Fig. 31.4 CAVA supports an iterative search process as described in the use case. This sequence shows several snapshots from the scenario where a clinician expands and refines an initial high-risk cohort using a mix of visual filters and patient similarity analytics. The end result is a targeted cohort of candidate patients for a new treatment regimen. (**a**) The sequence begins with a cohort overview showing age, gender, and diagnosis distributions. (**b**) Interactive visual filters are used to focus the analysis to narrower cohort. (**c**) Because the filtered group is too small, patient similarity analytics are requested to expand the cohort by retrieving additional clinically similar patients. The newly retrieved patients are visually integrated into the display for further analysis



Details

High risk based on hereditary profile

All patients in [High risk based on hereditary profile]

217 patients

Details

High risk based on hereditary profile

Gender: Male;

0 patients

Details

High risk based on hereditary profile

Diseases: Hypertension

128 patients

that having a system like INFUSE allows them to build a much richer picture of what kind of feature sets may lead to effective modeling. Without such a tool they would be restricted at evaluating one single algorithm at a time or, at best, restricting the comparison to a small number of features. Without such knowledge, the efficacy of the predictive models could be reduced.

31.4 Patient Stratification and Cohort Analysis

Patient stratification and cohort analysis are important techniques used in healthcare to study risk factors within population groups. The cohort study is a foundational tool that helps experts uncover correlations between specific risk metrics and the underlying attributes of individuals within the study population. Cohort studies are often performed prospectively using techniques that are statistically mature and powerful. However, the analytical process is often slow and expensive when collecting data prospectively. Retrospective analyses, which use previously collected data, are a possible alternative. Unfortunately, the use of retrospective studies has been relatively limited due to the historical difficulty in collecting and analyzing very large datasets. However, as more and more data become electronic, very large repositories suitable for retrospective cohort analysis are becoming increasingly common. These data warehouses can contain comprehensive historical observations of millions of people over time spans of many years. The increasing availability of such data helps overcome the fundamental limitations of the retrospective approach. In theory, domain experts can use these data to perform interactive, exploratory cohort studies without the overheads associated with prospective techniques. In practice, however, interactive cohort studies exploring large-scale retrospective data collections produce their own set of challenges. Data management, analysis, and summarization all become more difficult and typically lead to the use of more advanced technologies. Instead of relying on a spreadsheet and some basic statistics, users must also use technologies such as machine learning, data mining, and visualization tools to help make sense of the large scale of data they wish to examine.

31.4.1 Actionable Risk Stratification

A key step in providing personalized care is to segment the patient cohort into more homogeneous groups in terms of risk factors, so that a customized treatment plan can be constructed for each group. We term this process *Actionable Risk Stratification* because it goes beyond the traditional approaches of stratifying patients based on a single risk score. While that approach can effectively identify the group of high-risk patients to focus resources on, it does not provided insights into what are the most important risk factors to manage for these patients. Specifically, patients with the same (high) risk score may have incurred that high risk for different reasons (e.g., different comorbidities) and thus need to be managed differently.

A major challenge for actionable risk stratification is the heterogeneity of patients' clinical conditions. For example, CHF patients may have different comorbidities, such as diabetes, kidney diseases, or lung diseases. In different comorbidity groups, the medical features that contribute to the risk, or risk factors, are different. One way to perform patient stratification while taking into consideration of the most important factors is to construct a patient similarity network using techniques discussed earlier in this chapter, and then perform graph based clustering over this similarity network. However such an approach has the limitation that there is often inherent ambiguity in part of the network, due to the complexity of patients' conditions. As a result a purely data-driven approach would often lead to results that are unstable (i.e., different segmentation could emerge with slight perturbation of the attributes) and difficult to interpret.

One way to address the inherent ambiguity in data is to bring in prior knowledge from domain experts and literature. Such knowledge can be used to guide the data driven segmentation process such that the results conform with crucial clinical insights that have already been validated through extensive clinical studies, and are thus more interpretable and actionable. To this end we have developed an approach called RISGAL (RISk Group anALysis), which is a novel semi-supervised learning framework for data- and knowledge-driven patient risk group exploration [29]. The input of RISGAL is a graph with nodes as patients and edges as patient similarities, as well as a set of knowledge-driven risk factors or groups provided by domain experts or extracted from literature. The output is a set of patient risk groups that align with those provided risk factors. The approach was applied to a real-world electronic medical record database to stratify a set of patients with respect to their risk of CHF onset and was show to be able to identify both data- and knowledge-driven risk groups with rich clinical insights.

31.4.2 Healthcare Utilization Analysis and Hot Spotting

Another area where patient stratification has important applications is in healthcare utilization analysis. Utilization analysis based on observational healthcare data collected through normal course of care delivery and carried out in a systematic manner can be leveraged to improve care delivery in many ways. For example, through "hot spotting", we wish to identify patients, in a timely manner, who are heavy users of the system and their patterns of use, so that targeted intense intervention and follow up programs can be put in place to address their needs and change the existing, potentially ineffective, utilization pattern [4]. In anomaly detection, the goal is to identify utilization patterns that are unusual given patients' clinical characteristics, including both underutilization and overutilization. The former may indicate a gap in medical service that if left unaddressed could result in further deterioration of patient's condition leading to situations requiring more costly and less effective interventions. The latter incurs unnecessary cost and waste of precious healthcare resources that could have been directed towards cases in real need.

We have developed a novel framework for utilization analysis designed to address these needs. The first component of the framework is Utilization Profiling and Hot Spotting. Here we use a vector space model to represent patient utilization profiles, and apply advanced clustering techniques to identify dominant utilization groups within a given population. The second component of the framework is Contextual Anomaly Detection for Utilization. Here we developed a novel method for *contextual anomaly detection* designed to detect utilization anomalies while taken into consideration the patients' clinical characteristics. In this method we first build models trained from observational data to compute the expected utilization levels for each patient given his/her clinical and demographic characteristics. We then examine the difference between the expected and actual levels based on well-established statistical testing methods to identify anomalies. This utilization analysis framework was tested and evaluated using outpatient data for a population of 7,667 diabetes patients collected over a 1 year period, and was shown to be effective in identifying clinically meaningful instances for both hot spotting and anomaly detection [8].

31.4.3 Interactive, Visual Cohort Analysis

CAVA—a platform for Cohort Analysis via Visual Analytics—was designed to help clinical researchers work faster and more independently when performing retrospective cohort studies Zhang et al. [35]. Motivated by the needs of real-world analysts working in the healthcare domain, CAVA follows a novel system design centered around three primary types of artifacts: (1) cohorts, (2) views, and (3) analytics. Cohorts are CAVA's fundamental data construct and represent a set of people and their associated properties. Views are visualization components that graphically display a cohort and allow users to directly manipulate or refine the underlying cohort. Analytics are computational elements that create, expand, and/or alter the contents of a cohort. In this way, CAVA treats both Views and Analytics as functional components that operate on an input cohort and produce an output cohort. Building on this design principle, CAVA allows users to chain together complex sequences of steps that intermix both manual and machine-driven cohort manipulations.

31.4.4 Use Case: Iterative Cohort Analysis

The CAVA platform enables a wide range of cohort analysis workflows. As an example, suppose a clinician who has recently become aware of a new preventive technique that has been shown to help delay or prevent certain types of patients from developing heart disease. In particular, the treatment has been studied most in male hypertensive patients between 60 and 80 years of age. Due to limited resources and potential side effects, the clinician wants to focus this new treatment regimen on only those patients who are both (a) at high risk of developing the disease and (b) best fit the selection criteria for which the treatment is most effective. The clinician uses CAVA to find a cohort of candidates for the treatment following a usage pattern that we call iterative search.

To start, the physician selects a high-risk group from the cohort panel that has been generated by a risk stratification analytic. The user then drags and drops the cohort onto the demographic overview visualization icon. This results in the visualization shown in Fig. 31.4a, which displays linked views of age, gender, and diagnosis distributions. The user interactively selects various elements in the visualizations to explore how these three demographic criteria are correlated.

Next, the clinician interacts with the visualization to select and filter the age group in which the treatment has been studied: 60–80 years of age. By selecting the age range in the histogram and clicking the filter button, the user modifies the cohort to exclude those outside the specified range. The clinician then selects the men in the cohort and applies an additional filter. The result is shown in Fig. 31.4b. As a result of the filters, the initial cohort has been reduced to a group roughly one-third in size. However, the clinician presumes that there are likely additional patients—missing from the current cohort—who are clinically similar to the visualized patients and could benefit from the treatment even if they do not strictly meet the inclusion criteria. Therefore, the clinician decides to search for similar patients by dragging the current cohort from the active view to the Patient Similarity entry in the analytic panel. In response, CAVA binds the visualized cohort to the analytic and presents the user with a dialog box to gather the needed input parameters. In particular, the clinician indicates that she wants to retrieve enough similar patients to double the size of the cohort. After clicking OK, CAVA runs the analytic and updates the visualization with the newly expanded cohort.

The visualization now shows the additional similar patients, but the clinician is still not finished. Because the treatment was designed for patients with hypertension, she selects the hypertension subgroup in the visualization (as shown in Fig. 31.4c) and applies one last filter. The clinician has now used a combination of ad hoc filters and analytics to identify an initial set of candidate patients to target with the newly available treatment. Moreover, they have accomplished this without the help of a technology team to write SQL queries, run analytics, or produce reports.

31.5 Care Pathway Analytics

Extracting insights from temporal event sequences, such as mining frequent patterns, is an important challenge in healthcare. However, despite the availability of temporal data and the common desire to extract knowledge, mining patterns from temporal event sequences is still a fundamental challenge in data mining [14].

Frequent Sequence Mining (FSM) techniques have emerged in the data mining community to find sets of frequently occurring subsequences. However, these algorithms often have constraints that limit its applicability to real-world data.

First, they may not take into account the multiple levels of detail present in healthcare data. For example, ICD-9 diagnostic codes (which encode symptoms, causes, and signs of diseases using ICD-9 standards) are organized according to a meaningful hierarchy. In EMRs, temporal events are often recorded at a specific level-of-detail to record maximum information about an event's type. FSM techniques applied to data with a large dictionary of event types will often suffer from computational complexity. Perhaps even more of a fundamental problem is that patterns extracted from a specific level- of-detail may impair an interpretable overview of patterns for users.

A second issue is that FSM techniques ignore the temporal context associated with data, and instead focus on the pure sequentially of events. However, for medical scenarios, if a certain amount of time elapsed between events, the events should not be considered as part of the same sequence, even if events are technically sequential in the event log.

A third issue is concurrency. Many FSM algorithms suffer from pattern explosion when there are many concurrent events. This is particularly troubling for medical data, as many systems may record data in low-resolution precision, such as a day, and many events may occur on the same day.

A fourth issue is outcome. Many FSM algorithms are agnostic to the types of patterns mined. However, in healthcare data, analysts may not just need a list of patterns but instead how each of the patterns correlate to an outcome measure.

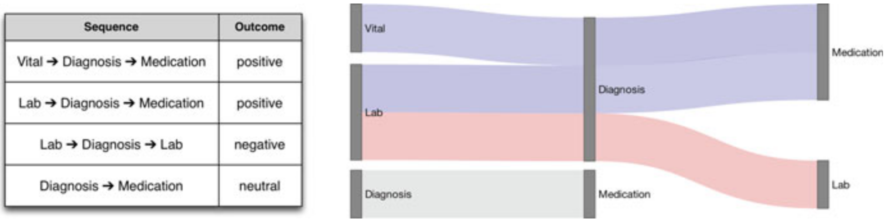
A recent system, Frequence [19], address these issues by featuring a novel frequent sequence mining algorithm to handle multiple levels-of-detail, temporal context, concurrency, and outcome analysis. Frequence also features a visual interface designed to support insights, and support exploration of patterns of the level-of-detail relevant to users.

31.5.1 Visual Representation of Frequent Patterns

In order to make the description of the system understandable, the characteristics of our visualization are illustrated in Fig. 31.5. In this example, the patterns are sequences of clinical events, and each patient has an outcome measure.

Events in the frequent sequences are represented as nodes, and event nodes that belong to the same sequence are connected by edges. The nodes and edges are positioned using a modified Sankey diagram layout [9].

Thus, in Frequence, subsequences are represented as individual edges. For instance, the simple pattern **Diagnosis** → **Medication**, is visualized as a **Diagnosis** node connected to a **Medication** node, as shown at the bottom of Fig. 31.5. Patterns that share similar subsequences, such as **Lab** → **Diagnosis** → **Medication** and **Lab** → **Diagnosis** → **Lab**, involve two edges from **Lab** to **Diagnosis** representing each



Sequence	Outcome
Vital ➔ Diagnosis ➔ Medication	positive
Lab ➔ Diagnosis ➔ Medication	positive
Lab ➔ Diagnosis ➔ Lab	negative
Diagnosis ➔ Medication	neutral

Fig. 31.5 An example of Frequence's visual encoding for a set of frequent patterns. Patterns are represented by a sequence of nodes (events) connected by edges (event subsequences). Patterns are colored according to their correlation with users' outcomes

subsequence. Thus, prominent subsequences also become visually prominent due to the thickness of the combined multiple edges.

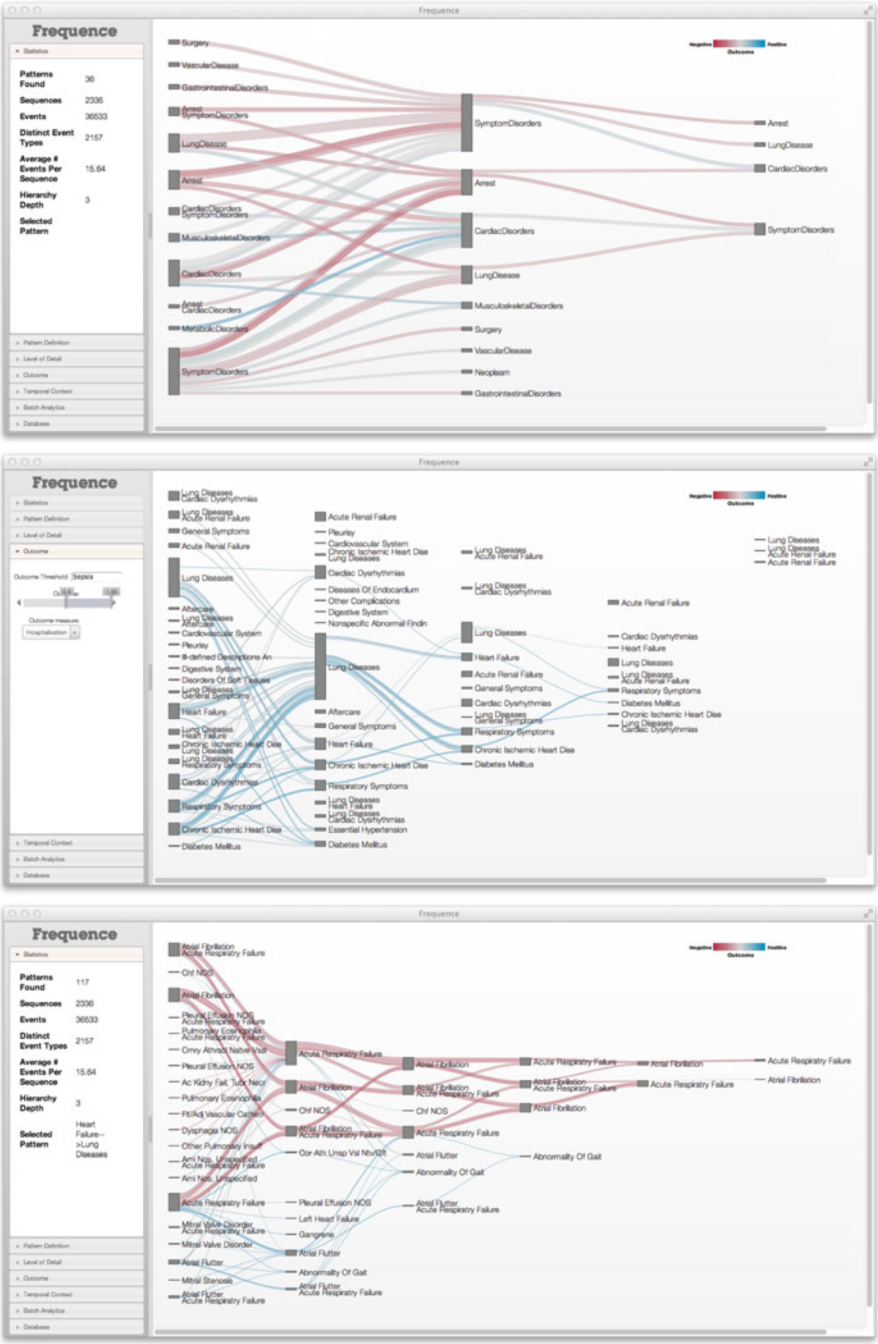
Of course, not all event subsequences are equal as some correlate to a positive outcome, whereas others correlate to a negative outcome, as determined by Frequence's outcome analytics. The visualization uses color to encode each pattern's association with an associated outcome. For this scenario, the patterns that occur more often with healthy patients are more blue. The patterns that occur more often with unhealthy patients are more red. The neutral patterns that appear common to both healthy and unhealthy patients are gray.

31.5.2 Use Case: Lung Disease and Sepsis

As an illustrative example, we briefly present a use case involving a team of clinical researchers interested in determining if there are particular patterns that lead to patients with lung disease developing sepsis, a potentially deadly medical condition. Additional details about this use case are presented in [19].

The institution used a set of 2,336 patients diagnosed with lung disease, each with longitudinal events of ICD-9 diagnostic codes. Of the patients with lung disease, 483 developed sepsis within 6 months of their diagnosis of lung disease, whereas 1,853 managed to not contract the condition.

At the top of Fig. 31.6, the coarsest patterns for all of the lung disease patients are shown. The clinician was particularly interested in cardiovascular complications, and noticed that the pattern **CardiacDisorders → SymptomDisorders** was common yet neutral (that is, this pattern was common to patients who did and did not end up contracting sepsis). After selecting this pattern in Frequence and filtering by cohort to see the matching patients, the finer level of detail (Level 1) allowed the clinician to see more detailed cardiac conditions, such as cardiac dysrhythmia and heart failure. Other complications, such as acute renal failure (which medical literature suggests is linked to developing sepsis), also appear. However, the clinician is interested in the patterns that led to patients not developing sepsis, and filtered to the positive patterns in the middle of Fig. 31.6. Surprised to see the pattern **HeartFailure**



Patterns Found 36
Sequences 2336
Events 36533
Distinct Event Types 2157
Average # Events Per Sequence 15.64
Hierarchy Depth 3
Selected Pattern

Outcome Threshold: Sepsis
Outcome measure: Hospitalization

Patterns Found 117
Sequences 2336
Events 36533
Distinct Event Types 2157
Average # Events Per Sequence 15.64
Hierarchy Depth 3
Selected Pattern Heart Failure->Lung Diseases

Fig. 31.6 The *top figure* shows an overview of the coarsest patterns in Frequence using the Lung Disease and Sepsis dataset. The *middle figure* shows the positive patterns at a finer level-of-detail for the cohort who matched the CardiacDisorders → SymptomDisorders sequence. The *bottom figure* shows the patterns at the finest level of detail, after selecting HeartFailure → LungDiseases

→ **LungDiseases**, the clinician filtered to the cohort that matched this pattern and pivoted to Level 2, as shown in the bottom of Fig. 31.6. The clinician immediately noticed that patterns that featured both Atrial Fibrillation and Acute Respiratory Failure are red, which is sensible, as medical literature suggests both are risk factors for sepsis. However, the clinician found it interesting that patterns beginning with Acute Respiratory Failure alone were not predictive of sepsis, but rather what happened next in the sequence was more predictive. From the Acute Respiratory Failure node in the first column of the visualization, the patterns diverge into red and blue, making it clear that what happens immediately after such Acute Respiratory Failure will likely determine if the patient will get sepsis or not.

31.6 Disease Modeling

Chronic diseases usually follow a long and slow progression. For example, Chronic Obstructive Pulmonary Disease (COPD) may take around 10 years to evolve from stage I (mild) to stage IV (very severe) [5]. It may also take 10 years for Congestive Heart Failure (CHF) progressing from stage I (mild) to stage IV (severe). Detection of such chronic diseases at its early stage is of key importance for effective treatment or intervention.

Disease Progression Modeling (DPM), which aims at modeling the entire progression procedure of a disease with computational technologies, is one important technique that can help realize disease early detection. Key challenges in developing DPM methodologies include: (1) Multiple Covariates. The progression of disease usually involves the evolution of many different types of covariates. In general it is not know which one or which group of variables are important. (2) Progression Heterogeneity. The patient disease conditions can progress differently for different individuals, and the patient records are not necessarily aligned. (3) Incomplete Records. The patient records are not complete, meaning that in most of the cases we are not able to get the patient records from the beginning stage of the disease to its end stage. (4) Irregular Visits. The patient only has medical records when he/she pays visit to medical facilities. Most of the times patients visits are at irregular time stamps due to various reasons. (5) Discrete Observation. Although the disease progression is a continuous time procedure, the patient records are only observed on certain discrete time stamps or intervals. (6) Limited Supervision. For most of the diseases we only have very limited knowledge on which observed events should belong to which disease stage.

As an initial step towards addressing these challenges, we developed an machine learning approach to infer probabilistic disease progression models from the longitudinal clinical findings of a cohort of patients who have developed, or are at risk developing such disease [26]. First of all, we use a Markov Jump Process to model the transition of disease stages/states, which implies (1) the progression is continuous-time; (2) the transition probability to the future state only relies the current state and the time span.

Second, we use the onset pattern of comorbidities to drive the transitions of the Markov Jump Process. Generally speaking, a comorbidity is a disease or syndrome that co-occurs with the target disease. For example, hypertension is a common comorbidity of diabetes and osteoporosis is a common comorbidity of COPD. Since the onset of a new comorbidity often signifies the exacerbation of the target disease, we use the onset pattern of multiple comorbidities to collectively capture the state transitions of the target disease. Finally, in order to infer the presence of the comorbidities from the observed clinical findings, we use a bipartite noisy- or Bayesian network [7, 20]. Simply speaking, given a set of comorbidities and a set of clinical findings, we assume an observed clinical finding was "activated" by the presence of any of the comorbidities with a certain activation probability. Such structure is especially well suited to our setting due to its flexibility in modeling sparse and noisy observations.

We validated our model on a data warehouse from a real-world longitudinal EMR database of 3,705 confirmed COPD patients over the course of 4 years. For each patient encounter ICD- 9 codes were recorded to indicate what medical conditions that patient had at that time point. Other information, such as drug prescription, lab test results, was also recorded. The results demonstrated that the proposed method can detect the episodes corresponding to different disease stage of every patient [26].

31.6.1 Visualizing Disease Progression

In order to better understand the progression of diseases, researchers can use tools like MatrixFlow. MatrixFlow is designed to help aid medical decision makers and researchers by making the subtle trends of disease progression more obvious. The goal is that by unearthing the hidden patterns in patient health records, emerging health risks may become more discoverable and earlier diagnoses of diseases can occur so clinicians and patients can proactively develop preventative strategies to reduce negative future outcomes.

The analytics work by extracting clinical event sequences from patient EMR data and then constructing a temporal network of co-occurring events to model the relationships between events as a disease progresses over time. The patterns in the evolution of the disease are then revealed in our interactive visualization as a temporal flow of matrices, MatrixFlow. MatrixFlow provides several interactive features for analysis: (1) one can sort the events based on the similarity in order to accentuate underlying cluster patterns among those events; (2) one can compare co-occurrence events over time and across cohorts through additional line graph visualization.

31.6.2 Clinical Event Networks

This work aims at discovering meaningful patterns from clinical event sequences of patients. Clinical event sequences are simply a series of time-stamped events from a patient's medical record, such as disease diagnoses, patient symptoms, lab results, and medication orders. However, what if researchers are interested in determining the co-occurrence of event—that is, when events simultaneously occur. Co-occurrence can be modeled by creating a network of clinical events, where events are nodes, and co-occurring events are connected by an edge.

Instead of using a traditional node-link diagram, MatrixFlow [18] relies on its namesake visualization: the adjacency matrix. In matrix visualizations, the columns and rows represent the nodes of the network, whereas each cell in a matrix represents the edge between the two nodes.

31.6.3 Use Case: Heart Failure

One motivating example is the clinical complexity and heterogeneity of heart failure (HF). HF has posed challenges to developing standardized criteria for its diagnosis. The Framingham HF criteria, originally published in 1971, were based on clinical data acquired in the 1950s and 1960s. In that study, two or more major criteria or one major and two or more minor criteria are used as the diagnosis criteria for HF. The challenges for making the correct HF diagnosis earlier are (1) how to correlate the sparse signals of a single patient across time and encounters, and (2) how to leverage historical data of other similar patients to identify the emerging pattern earlier.

We illustrate the capabilities of MatrixFlow with a dataset of over 50,625 patients. A total of 4,644 incident HF cases were identified between 2003 and 2010. Up to ten control patients were selected for each case. Controls were clinic-matched, sex-matched, and age-matched to the corresponding case but did not meet operational criteria for HF on or before the corresponding case's diagnosis date. Note that two different cases can share common controls, in this design. For this study, we extracted the clinical notes portion of the EMRs for 4,644 case patients and for 45,981 control patients. Additionally, we have 1,200 confirmed HRrEF (reduced ejection fraction) and 1,615 confirmed HFpEF (preserved ejection fraction) cases, and the rest are HF cases without a confirmed subtype.

Figure 31.7a shows the evolution of co-occurrence matrices of positive Framingham symptoms in the HFrEF patients, where patients are aligned by their diagnosis date. Each matrix displays co-occurrence events in a 3-month window. The rightmost matrix corresponds to the window right before diagnosis and the leftmost one the window 15–18 months before the diagnosis. From left to right as time evolves, it is possible to observe the percentage of patients having co-occurring Framingham symptoms is increasing, which confirms with the degrading clinical status of those patients. Notably, as patients gets closer to HF diagnosis, multiple Framingham symptoms starts to appear more frequently. A similar temporal pattern is observed in HFpEF patients Fig. 31.7b, which seems to suggest that despite the pathophysiological differences, both HF types seem to develop the same co-occurrence patterns on Framingham symptoms. On the other hand, control groups

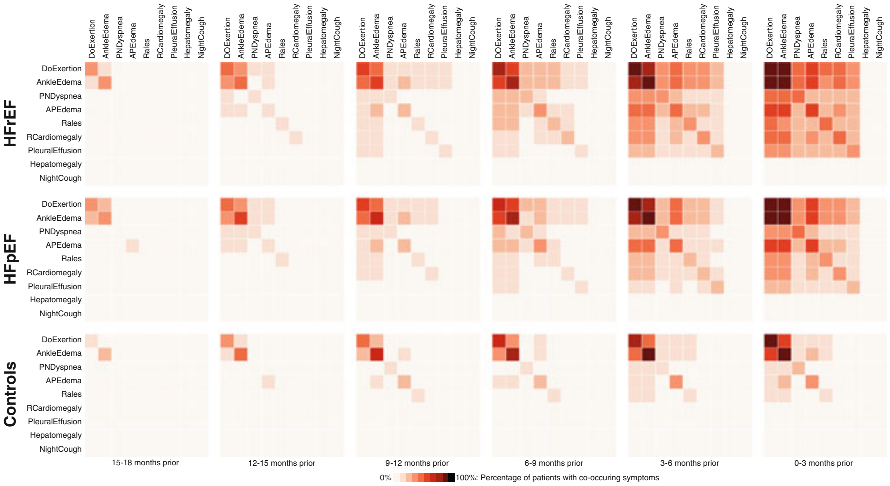


Fig. 31.7 The temporal evolution of the Framingham symptoms in MatrixFlow. (**a**) The *top row* of matrices represents the patterns the HFrEF patient cohort. (**b**) The *middle row* represents the HFpEF patient cohort. (**c**) The *bottom row* represents the Controls cohort

Fig. 31.7c have much less obvious patterns, except a slight increase of prevalence on common symptoms like DOExertion and AnkleEdema, presumably due to the normal aging process.

31.7 Conclusions

Healthcare has undergone a tremendous growth in the use of EMR systems to capture patient disease and treatment histories. This and other rich observational data being captured in the healthcare system provide the foundational source material for realizing the vision of Learning Health Systems. However, to truly realize this vision, advanced data driven analytics and visualization methodologies and systems need to be developed in order to convert the source material into meaningful insights. In this chapter we described some initial progresses we have made in applying advanced analytics to derive insights to support smarter, more personalized care, and the journey continues. We are continuing to develop cutting edge innovations that will take us to the next level. One direction we are exploring is the integration of health and social programs. Specifically, we are working on developing comprehensive risk assessment models to better predictive risk in a cross domain environment, linking together physical health with mental health, behavioral factors, and overall quality of life, in order to drive successful integrated care. Another direction we are focusing on is the better understanding of the drivers of risk through disease modeling. Here we have efforts underway to better model phenotype representation from complex data, and to develop machine learning approaches to derive disease progression models. Finally, we are working on developing a *Visual*

Analytics Workbench, where advanced analytics can be tightly integrated with interactive visualizations tools to support dynamic, comprehensive and efficient data-driven hypothesis generating and testing. Our vision here is to provide a powerful tool that can be leveraged by researchers everywhere to speed up the development of data-driven analytics that can lead to better deliver of care at lower cost.

References

1. Alexander GC, Stafford RS. Does comparative effectiveness have a comparative edge? JAMA. 2009;301:2488–90.
2. Berwick DM. Disseminating innovations in healthcare. JAMA. 2003;289:1969–75.
3. Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. AMIA Annu Symp Proc. 2010;2010:192–6.
4. Gawande A. The hot spotters. New Yorker, Jan 2011.
5. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of COPD. 2014. http://www.goldcopd.org/guidelines-global-strategy-for-diagnosis-management.html. Accessed 21 Apr 2015.
6. Gotz D, Starvropoulos H, Sun J, Wang F. ICDA: a platform for intelligent care delivery analytics. Am Med Inform Assoc Annu Symp AMIA. 2012;2012:264–73.
7. Halpern Y, Sontag D. Unsupervised learning of noisy-or bayesian networks. In: Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence (UAI2013). Bellevue, WA, USA. 2013. p. 272–81. arXiv:1309.6834 [cs.LG].
8. Hu J, Wang F, Sun J, Sorrentino R, Ebadollahi S. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. Am Med Inform Assoc Annu Symp (AMIA 2012). 2012;2012:360–9.
9. Krause J, Perer A, Bertini E. INFUSE: interactive feature selection for predictive modelling of high dimensional data. Paris: IEEE Visual Analytics Science and Technology (VAST 2014); 2014.
10. Lenfant C. Clinical research to clinical practice – lost in translation. N Engl J Med. 2003;349: 868–74.
11. LHS. Institute of Medicine Report: best care at lower cost: the path to continuously learning health care in America, released on 6 Sept 2012. 2013. http://www.iom.edu/Reports/2012/Best-Care-at-Lower-Cost-The-Path-to-Continuously-Learning-Health-Care-in-America.aspx.
12. Luo D, Wang F, Sun J, Markatou M, Hu J, Ebadollahi S. SOR: scalable orthogonal regression for non redundant feature selection and its healthcare applications. SIAM Data Mining. 2012. http://www.research.ibm.com/healthcare/papers/sorSDM2012.pdf.
13. Markatou M, Kuruppumullage Don P, Hu J, Wang F, Sun J, Sorrentino R, Ebadollahi S. Case-based reasoning in comparative effectiveness research. IBM J Dev Res. 2012;56(5):468–79.
14. Mitsa T, editor. Temporal data mining. 1st ed. Boca Raton: Chapman & Hall/CRC; 2010.
15. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J Biomed Inform. 2014;48:160–70.
16. Partners Healthcare. i2b2. 2014. https://www.i2b2.org/.
17. Perer A, Gotz D. Data driven exploration of care plans for patients. Paris: ACM CHI; 2013.
18. Perer A, Sun J. MatrixFlow: temporal network visual analytics to track symptom evolution during disease progression. Am Med Inform Assoc Annu Symp (AMIA 2012). 2012;2012: 716–25.
19. Perer A, Wang F. Frequence: interactive mining and visualization of temporal frequent event sequences. In: IUI '14 proceedings of the 19th international conference on intelligent user interfaces. New York: ACM; 2014. doi:10.1145/2557500.2557508.

20. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. Lifelines: using visualization to enhance navigation and analysis of patient records. In American Medical Informatics Association Annual Symposium (AMIA), AMIA 1998 (1998), 7680.
21. Shwe MA, Middleton B, Heckerman D, Henrion M, Horvitz E, Lehmann H, Cooper G. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. Methods Inf Med. 1991;30:241–55.
22. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S, Steinhubl SE, Daar Z, Stewart WF. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA. 2012;2012:901–10.
23. Sun J, Sow DM, Hu J, Ebadollahi S. A system for mining temporal physiological data streams for advanced prognostic decision support. In: IEEE international conference on data mining. 2010. p. 1061–66. http://www.research.ibm.com/healthcare/papers/05694085.pdf.
24. Sun J, Sow DM, Hu J, Ebadollahi S. Localized supervised metric learning on temporal physiological data. In: International conference on pattern recognition. 2010. p. 4149–52. http://www.research.ibm.com/healthcare/papers/05597728.pdf.
25. Tracy CS, Dantas G, Upshur R. Evidence- based medicine in primary care: qualitative study of family physicians. BMC Fam Pract. 2003;4(1):6.
26. Wang X, Sontag D, Wang F. Unsupervised learning of disease progression models. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2014. p. 85–94.
27. Wang F, Sun J, Hu J, Ebadollahi S. iMet: interactive metric learning in healthcare applications. In: SIAM Data Mining Conference. 2011. pp. 944–55. http://www.research.ibm.com/healthcare/papers/304.pdf.
28. Wang F, Sun J, Ebadollahi S. Integrating distance metrics learned from multiple experts and its application in inter-patient similarity assessment. In: SIAM Data Mining Conference. 2011. p. 59–70. http://www.research.ibm.com/healthcare/papers/113.pdf.
29. Wang X, Wang F, Wang J, Qian B, Hu J. Exploring patient risk groups with incomplete knowledge. 2013 IEEE 13th international conference on data mining (ICDM). New York: IEEE; 2013. p. 1223–28.
30. Wang F, Zhang C. Feature extraction by maximizing the average neighborhood margin. In: Computer Vision and Pattern Recognition, New York: IEEE; 2007. p. 1–8.
31. Wang F, Zhang C. Label propagation through linear neighborhoods. In: Proceedings of the 23rd international conference on machine learning, Pittsburgh, 2006, p. 985–92. http://www.autonlab.org/icml_documents/camera-ready/124_Label_Propagation_th.pdf.
32. WDA. What is Watson? 2014. http://www.ibm.com/smarterplanet/us/en/ibmwatson/discovery-advisor.html.
33. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. J Am Med Inform Assoc. 2013;20(5):954–61.
34. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668–72.
35. Zhang Z, Gotz D, Perer A. Iterative cohort analysis and exploration. Journal of Information Visualization, March 19, 2014. doi: 10.1177/1473871614526077. http://ivi.sagepub.com/content/early/2014/03/19/1473871614526077.abstract
36. Zhang P, Wang F, Hu J, Sorrentino R. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. Am Med Inform Assoc (AMIA) Jt Summit Transl Sci Transl Bioinforma (TBI). 2014;2014:132–6.
37. Zhou J, Wang F, Hu J, Ye J. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2014. p. 135–44.